

A black metal spiral staircase with a brick wall background. The staircase is the central focus, winding upwards from the bottom right towards the top left. The brick wall is made of reddish-brown bricks with white mortar. The lighting is bright, creating strong shadows and highlights on the metal and bricks.

# Basic Statistics for Business & Economics

Ninth Edition

**Mc  
Graw  
Hill**  
Education

LIND

MARCHAL

WATHEN

Basic Statistics for  
**BUSINESS &  
ECONOMICS**

# The McGraw-Hill/Irwin Series in Operations and Decision Sciences

## SUPPLY CHAIN MANAGEMENT

Benton

### **Purchasing and Supply Chain Management**

*Third Edition*

Bowersox, Closs, Cooper, and Bowersox

### **Supply Chain Logistics Management**

*Fourth Edition*

Burt, Petcavage, and Pinkerton

### **Supply Management**

*Eighth Edition*

Johnson, Leenders, and Flynn

### **Purchasing and Supply Management**

*Fifteenth Edition*

Simchi-Levi, Kaminsky, and Simchi-Levi

### **Designing and Managing the Supply Chain: Concepts, Strategies, Case Studies**

*Third Edition*

## PROJECT MANAGEMENT

Brown and Hyer

### **Managing Projects: A Team-Based Approach**

*First Edition*

Larson and Gray

### **Project Management: The Managerial Process**

*Sixth Edition*

## SERVICE OPERATIONS MANAGEMENT

Fitzsimmons and Fitzsimmons

### **Service Management: Operations, Strategy, Information Technology**

*Ninth Edition*

## MANAGEMENT SCIENCE

Hillier and Hillier

### **Introduction to Management Science: A Modeling and Case Studies Approach with Spreadsheets**

*Sixth Edition*

Stevenson and Ozgur

### **Introduction to Management Science with Spreadsheets**

*First Edition*

## MANUFACTURING CONTROL SYSTEMS

Jacobs, Berry, Whybark, and Vollmann

### **Manufacturing Planning & Control for Supply Chain Management**

*Sixth Edition*

## BUSINESS RESEARCH METHODS

Cooper and Schindler

### **Business Research Methods**

*Twelfth Edition*

## BUSINESS FORECASTING

Wilson, Keating, and John Galt Solutions, Inc.

### **Business Forecasting**

*Seventh Edition*

## LINEAR STATISTICS AND REGRESSION

Kutner, Nachtsheim, and Neter

### **Applied Linear Regression Models**

*Fourth Edition*

## BUSINESS SYSTEMS DYNAMICS

Sterman

### **Business Dynamics: Systems Thinking and Modeling for a Complex World**

*First Edition*

## OPERATIONS MANAGEMENT

Cachon and Terwiesch

### **Matching Supply with Demand: An Introduction to Operations Management**

*Fourth Edition*

Finch

### **Interactive Models for Operations and Supply Chain Management**

*First Edition*

Jacobs and Chase

### **Operations and Supply Chain Management**

*Fifteenth Edition*

Jacobs and Chase

### **Operations and Supply Chain Management: The Core**

*Fourth Edition*

Jacobs and Whybark

### **Why ERP? A Primer on SAP Implementation**

*First Edition*

Schroeder, Goldstein, and

Rungtusanatham

### **Operations Management in the Supply Chain: Decisions and Cases**

*Seventh Edition*

Stevenson

### **Operations Management**

*Twelfth Edition*

Swink, Melnyk, Cooper, and Hartley

### **Managing Operations across the Supply Chain**

*Third Edition*

## PRODUCT DESIGN

Ulrich and Eppinger

### **Product Design and Development**

*Sixth Edition*

## BUSINESS MATH

Slater and Wittry

### **Math for Business and Finance: An Algebraic Approach**

*Second Edition*

Slater and Wittry

### **Practical Business Math Procedures**

*Twelfth Edition*

## BUSINESS STATISTICS

Bowerman, O'Connell, and Murphree

### **Business Statistics in Practice**

*Eighth Edition*

Bowerman, O'Connell, Murphree, and Orris

### **Essentials of Business Statistics**

*Fifth Edition*

Doane and Seward

### **Applied Statistics in Business and Economics**

*Fifth Edition*

Lind, Marchal, and Wathen

### **Basic Statistics for Business and Economics**

*Ninth Edition*

Lind, Marchal, and Wathen

### **Statistical Techniques in Business and Economics**

*Seventeenth Edition*

Jaggia and Kelly

### **Business Statistics: Communicating with Numbers**

*Second Edition*

Jaggia and Kelly

### **Essentials of Business Statistics: Communicating with Numbers**

*First Edition*

Basic Statistics for  
**BUSINESS &  
ECONOMICS**

**NINTH EDITION**

**DOUGLAS A. LIND**

*Coastal Carolina University and The University of Toledo*

**WILLIAM G. MARCHAL**

*The University of Toledo*

**SAMUEL A. WATHEN**

*Coastal Carolina University*





## BASIC STATISTICS FOR BUSINESS AND ECONOMICS, NINTH EDITION

Published by McGraw-Hill Education, 2 Penn Plaza, New York, NY 10121. Copyright © 2019 by McGraw-Hill Education. All rights reserved. Printed in the United States of America. Previous editions © 2013, 2011, and 2008. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written consent of McGraw-Hill Education, including, but not limited to, in any network or other electronic storage or transmission, or broadcast for distance learning.

Some ancillaries, including electronic and print components, may not be available to customers outside the United States.

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 LWI 21 20 19 18

ISBN 978-1-260-18750-2

MHID 1-260-18750-0

Portfolio Manager: *Noelle Bathurst*

Product Developers: *Michele Janicek / Ryan McAndrews*

Marketing Manager: *Harper Christopher*

Content Project Manager: *Lori Koettters*

Buyer: *Sandy Ludovissy*

Design: *Matt Backhaus*

Content Licensing Specialist: *Ann Marie Jannette*

Cover Image: ©*Ingram Publishing / SuperStock*

Compositor: *Aptara®, Inc.*

All credits appearing on page or at the end of the book are considered to be an extension of the copyright page.

### Library of Congress Cataloging-in-Publication Data

Names: Lind, Douglas A., author. | Marchal, William G., author. | Wathen, Samuel Adam. author.

Title: Basic statistics for business and economics / Douglas A. Lind, Coastal Carolina University and The University of Toledo, William G. Marchal, The University of Toledo, Samuel A. Wathen, Coastal Carolina University.

Description: Ninth edition. | New York, NY : McGraw-Hill Education, [2019]

Identifiers: LCCN 2017034976 | ISBN 9781260187502 (alk. paper)

Subjects: LCSH: Social sciences—Statistical methods. |

Economics—Statistical methods. | Industrial management—Statistical methods.

Classification: LCC HA29 .L75 2019 | DDC 519.5—dc23 LC record available at

<https://lccn.loc.gov/2017034976>

The Internet addresses listed in the text were accurate at the time of publication. The inclusion of a website does not indicate an endorsement by the authors or McGraw-Hill Education, and McGraw-Hill Education does not guarantee the accuracy of the information presented at these sites.

## **DEDICATION**

*To Jane, my wife and best friend, and our sons, their wives, and our grandchildren: Mike and Sue (Steve and Courtney), Steve and Kathryn (Kennedy, Jake, and Brady), and Mark and Sarah (Jared, Drew, and Nate).*

Douglas A. Lind

*To Oscar Sambath Marchal, Julian Irving Horowitz, Cecilia Marchal Nicholson, and Andrea.*

William G. Marchal

*To my wonderful family: Barb, Hannah, and Isaac.*

Samuel A. Wathen

# A NOTE FROM THE AUTHORS

Over the years, we received many compliments on this text and understand that it's a favorite among students. We accept that as the highest compliment and continue to work very hard to maintain that status.

The objective of *Basic Statistics for Business and Economics* is to provide students majoring in management, marketing, finance, accounting, economics, and other fields of business administration with an introductory survey of descriptive and inferential statistics. To illustrate the application of statistics, we use many examples and exercises that focus on business applications, but also relate to the current world of the college student. A previous course in statistics is not necessary, and the mathematical requirement is first-year algebra.

In this text, we show beginning students every step needed to be successful in a basic statistics course. This step-by-step approach enhances performance, accelerates preparedness, and significantly improves motivation. Understanding the concepts, seeing and doing plenty of examples and exercises, and comprehending the application of statistical methods in business and economics are the focus of this book.

The first edition of this text was published in 1967. At that time, locating relevant business data was difficult. That has changed! Today, locating data is not a problem. The number of items you purchase at the grocery store is automatically recorded at the checkout counter. Phone companies track the time of our calls, the length of calls, and the identity of the person called. Credit card companies maintain information on the number, time and date, and amount of our purchases. Medical devices automatically monitor our heart rate, blood pressure, and temperature from remote locations. A large amount of business information is recorded and reported almost instantly. CNN, *USA Today*, and MSNBC, for example, all have websites that track stock prices in real time.

Today, the practice of data analytics is widely applied to “big data.” The practice of data analytics requires skills and knowledge in several areas. Computer skills are needed to process large volumes of information. Analytical skills are needed to evaluate, summarize, organize, and analyze the information. Critical thinking skills are needed to interpret and communicate the results of processing the information.

Our text supports the development of basic data analytical skills. In this edition, we added a new section at the end of each chapter called Data Analytics. As you work through the text, this section provides the instructor and student with opportunities to apply statistical knowledge and statistical software to explore several business environments. Interpretation of the analytical results is an integral part of these exercises.

A variety of statistical software is available to complement our text. Microsoft Excel includes an add-in with many statistical analyses. MegaStat is an add-in available for Microsoft Excel. Minitab and JMP are stand-alone statistical software available to download for either PC or Mac computers. In our text, Microsoft Excel, Minitab, and MegaStat are used to illustrate statistical software analyses. When a software application is presented, the software commands for the application are available in Appendix C. We use screen captures within the chapters, so the student becomes familiar with the nature of the software output.

Because of the availability of computers and software, it is no longer necessary to dwell on calculations. We have replaced many of the calculation examples with interpretative ones, to assist the student in understanding and interpreting the statistical results. In addition, we place more emphasis on the conceptual nature of the statistical topics. While making these changes, we still continue to present, as best we can, the key concepts, along with supporting interesting and relevant examples.

## WHAT'S NEW IN THE NINTH EDITION?

We have made many changes to examples and exercises throughout the text. The section on “Enhancements” to our text details them. There are two major changes to the text. First, the chapters have been reorganized so that each section corresponds to a learning objective. The learning objectives have been revised.

The second major change responds to user interest in the area of data analytics. Our approach is to provide instructors and students with the opportunity to combine statistical knowledge, computer and statistical software skills, and interpretative and critical thinking skills. A set of new and revised exercises is included at the end of each chapter in a section titled “Data Analytics.”

In these sections, exercises refer to three data sets. The North Valley Real Estate sales data set lists 105 homes currently on the market. The Lincolnville School District bus data list information on 80 buses in the school district’s bus fleet. The authors designed these data so that students will be able to use statistical software to explore the data and find realistic relationships in the variables. The Baseball Statistics for the 2016 season is updated from the previous edition.

The intent of the exercises is to provide the basis of a continuing case analysis. We suggest that instructors select one of the data sets and assign the corresponding exercises as each chapter is completed. Instructor feedback regarding student performance is important. Students should retain a copy of each chapter’s results and interpretations to develop a portfolio of discoveries and findings. These will be helpful as students progress through the course and use new statistical techniques to further explore the data. The ideal ending for these continuing data analytics exercises is a comprehensive report based on the analytical findings.

We know that working with a statistics class to develop a very basic competence in data analytics is challenging. Instructors will be teaching statistics. In addition, instructors will be faced with choosing statistical software and supporting students in developing or enhancing their computer skills. Finally, instructors will need to assess student performance based on assignments that include both statistical and written components. Using a mentoring approach may be helpful.

We hope that you and your students find this new feature interesting and engaging.



# HOW ARE CHAPTERS ORGANIZED TO ENGAGE STUDENTS AND PROMOTE LEARNING?

## Chapter Learning Objectives

Each chapter begins with a set of learning objectives designed to provide focus for the chapter and motivate student learning. These objectives, located in the margins next to the topic, indicate what the student should be able to do after completing each section in the chapter.

▲ **MERRILL LYNCH** recently completed a study of online investment portfolios for a sample of clients. For the 70 participants in the study, organize these data into a frequency distribution. (See Exercise 43 and **LO2-3**.)

### LEARNING OBJECTIVES

When you have completed this chapter, you will be able to:

- LO2-1** Summarize qualitative variables with frequency and relative frequency tables.
- LO2-2** Display a frequency table using a bar or pie chart.
- LO2-3** Summarize quantitative variables with frequency and relative frequency distributions.
- LO2-4** Display a frequency distribution using a histogram or frequency polygon.

## Chapter Opening Exercise

A representative exercise opens the chapter and shows how the chapter content can be applied to a real-world situation.

## Introduction to the Topic

Each chapter starts with a review of the important concepts of the previous chapter and provides a link to the material in the current chapter. This step-by-step approach increases comprehension by providing continuity across the concepts.

### INTRODUCTION

The United States automobile retailing industry is highly competitive. It is dominated by megadealerships that own and operate 50 or more franchises, employ over 10,000 people, and generate several billion dollars in annual sales. Many of the top dealerships are publicly owned, with shares traded on the New York Stock Exchange or NASDAQ. In 2017, the largest megadealership was AutoNation (ticker symbol AN), followed by Penske Auto Group (PAG), Group 1 Automotive, Inc. (ticker symbol GPI), and the privately owned Van Tuyl Group.



©Darren Brode/Shutterstock

These large corporations use statistics and analytics to summarize and analyze data and information to support their decisions. As an example, we will look at the Applewood Auto Group. It owns four dealerships and sells a wide range of vehicles. These include the popular Korean brands Kia and Hyundai, BMW and Volvo sedans and luxury SUVs, and a full line of Ford and Chevrolet cars and trucks.

Ms. Kathryn Ball is a member of the senior management team at Applewood Auto Group, which has its corporate offices adjacent to Kane

## Example/Solution

After important concepts are introduced, a solved example is given. This example provides a how-to illustration and shows a relevant business application that helps students answer the question, “How can I apply this concept?”

### EXAMPLE

Ms. Kathryn Ball of the Applewood Auto Group wants to summarize the quantitative variable profit with a frequency distribution and display the distribution with charts and graphs. With this information, Ms. Ball can easily answer the following questions: What is the typical profit on each sale? What is the largest or maximum profit on any sale? What is the smallest or minimum profit on any sale? Around what value do the profits tend to cluster?

### SOLUTION

To begin, we show the profits for each of the 180 vehicle sales listed in Table 2–4. This information is called raw or ungrouped data because it is simply a listing

**TABLE 2–4** Profit on Vehicles Sold Last Month by the Applewood Auto Group

|         |         |         |        |        |         |         |         |         | Maximum |
|---------|---------|---------|--------|--------|---------|---------|---------|---------|---------|
| \$1,387 | \$2,148 | \$2,201 | \$ 963 | \$ 820 | \$2,230 | \$3,043 | \$2,584 | \$2,370 |         |
| 1,754   | 2,207   | 996     | 1,298  | 1,266  | 2,341   | 1,059   | 2,666   | 2,637   |         |
| 1,817   | 2,252   | 2,813   | 1,410  | 1,741  | 3,292   | 1,674   | 2,991   | 1,426   |         |

## Self-Reviews

Self-Reviews are interspersed throughout each chapter and follow Example/Solution sections. They help students monitor their progress and provide immediate reinforcement for that particular technique. Answers are in Appendix E.

### SELF-REVIEW 2-5



The hourly wages of the 15 employees of Matt’s Tire and Auto Repair are organized into the following table.

| Hourly Wages    | Number of Employees |
|-----------------|---------------------|
| \$ 8 up to \$10 | 3                   |
| 10 up to 12     | 7                   |
| 12 up to 14     | 4                   |
| 14 up to 16     | 1                   |

- (a) What is the table called?
- (b) Develop a cumulative frequency distribution and portray the distribution in a cumulative frequency polygon.
- (c) On the basis of the cumulative frequency polygon, how many employees earn less than \$11 per hour?

## Statistics in Action

Statistics in Action articles are scattered throughout the text, usually about two per chapter. They provide unique, interesting applications and historical insights in the field of statistics.

### STATISTICS IN ACTION

Florence Nightingale is known as the founder of the nursing profession. However, she also saved many lives by using statistical analysis. When she encountered an unsanitary condition or an undersupplied hospital, she improved the conditions and then

## Definitions

Definitions of new terms or terms unique to the study of statistics are set apart from the text and highlighted for easy reference and review. They also appear in the Glossary at the end of the book.

**HISTOGRAM** A graph in which the classes are marked on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are represented by the heights of the bars, and the bars are drawn adjacent to each other.

## Formulas

Formulas that are used for the first time are boxed and numbered for reference. In addition, key formulas are listed in the back of the text as a reference.

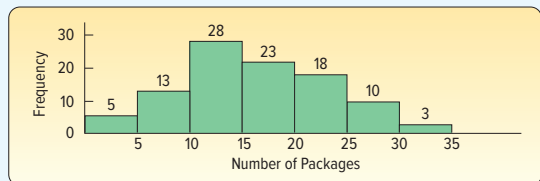
**CLASSICAL PROBABILITY**  $\text{Probability of an event} = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$  (5-1)

## Exercises

Exercises are included after sections within the chapter and at the end of the chapter. Section exercises cover the material studied in the section. Many exercises have data files available to import into statistical software. They are indicated with the FILE icon. Answers to the odd-numbered exercises are in Appendix D.

### EXERCISES

15. Molly's Candle Shop has several retail stores in the coastal areas of North and South Carolina. Many of Molly's customers ask her to ship their purchases. The following chart shows the number of packages shipped per day for the last 100 days. For example, the first class shows that there were 5 days when the number of packages shipped was 0 up to 5.



## Computer Output

The text includes many software examples, using Excel, MegaStat, and Minitab. The software results are illustrated in the chapters. Instructions for a particular software example are in Appendix C.

| APPLEWOOD AUTO GROUP |     |         |           |              |          |   |                    |           |
|----------------------|-----|---------|-----------|--------------|----------|---|--------------------|-----------|
|                      | A   | B       | C         | D            | E        | F | G                  | H         |
| 1                    | Age | Profit  | Location  | Vehicle-Type | Previous |   | Profit             |           |
| 2                    | 21  | \$1,387 | Tionesta  | Sedan        | 0        |   |                    |           |
| 3                    | 23  | \$1,754 | Sheffield | SUV          | 1        |   | Mean               | 1843.17   |
| 4                    | 24  | \$1,817 | Sheffield | Hybrid       | 1        |   | Standard Error     | 47.97     |
| 5                    | 25  | \$1,040 | Sheffield | Compact      | 0        |   | Median             | 1882.50   |
| 6                    | 26  | \$1,273 | Kane      | Sedan        | 1        |   | Mode               | 1915.00   |
| 7                    | 27  | \$1,529 | Sheffield | Sedan        | 1        |   | Standard Deviation | 643.63    |
| 8                    | 27  | \$3,082 | Kane      | Truck        | 0        |   | Sample Variance    | 414256.61 |
| 9                    | 28  | \$1,951 | Kane      | SUV          | 1        |   | Kurtosis           | -0.22     |
| 10                   | 28  | \$2,692 | Tionesta  | Compact      | 0        |   | Skewness           | -0.24     |
| 11                   | 29  | \$1,342 | Kane      | Sedan        | 2        |   | Range              | 2998      |
| 12                   | 29  | \$1,206 | Sheffield | Sedan        | 0        |   | Minimum            | 294       |
| 13                   | 30  | \$443   | Kane      | Sedan        | 3        |   | Maximum            | 3292      |
| 14                   | 30  | \$1,621 | Sheffield | Truck        | 1        |   | Sum                | 331770    |
| 15                   | 30  | \$754   | Olean     | Sedan        | 2        |   | Count              | 180       |

Source: Microsoft Excel

# HOW DOES THIS TEXT REINFORCE STUDENT LEARNING?

## BY CHAPTER

### Chapter Summary

Each chapter contains a brief summary of the chapter material, including vocabulary, definitions, and critical formulas.

| CHAPTER SUMMARY  |       |
|--|-------|
| I. A measure of location is a value used to describe the central tendency of a set of data.                    |       |
| A. The arithmetic mean is the most widely reported measure of location.  |       |
| 1. It is calculated by adding the values of the observations and dividing by the total number of observations. |       |
| a. The formula for the population mean of ungrouped or raw data is   |       |
| $\mu = \frac{\sum x}{N}$   | (3-1) |
| b. The formula for the sample mean is  |       |
| $\bar{x} = \frac{\sum x}{n}$   | (3-2) |
| 2. The major characteristics of the arithmetic mean are:   |       |

### Pronunciation Key

This section lists the mathematical symbol, its meaning, and how to pronounce it. We believe this will help the student retain the meaning of the symbol and generally enhance course communications.

| PRONUNCIATION KEY |                               |                      |
|-------------------|-------------------------------|----------------------|
| SYMBOL            | MEANING                       | PRONUNCIATION        |
| $\mu$             | Population mean               | <i>mu</i>            |
| $\Sigma$          | Operation of adding           | <i>sigma</i>         |
| $\Sigma x$        | Adding a group of values      | <i>sigma x</i>       |
| $\bar{x}$         | Sample mean                   | <i>x bar</i>         |
| $\bar{x}_w$       | Weighted mean                 | <i>x bar sub w</i>   |
| $\sigma^2$        | Population variance           | <i>sigma squared</i> |
| $\sigma$          | Population standard deviation | <i>sigma</i>         |

### Chapter Exercises

Generally, the end-of-chapter exercises are the most challenging and integrate the chapter concepts. The answers and worked-out solutions for all odd-numbered exercises are in Appendix D at the end of the text. Many exercises are noted with a data file icon in the margin. For these exercises, there are data files in Excel format located in Connect. These files help students use statistical software to solve the exercises.

| CHAPTER EXERCISES  |  |                         |    |                                |     |          |    |           |    |
|--|--|-------------------------|----|--------------------------------|-----|----------|----|-----------|----|
| 23. Describe the similarities and differences of qualitative and quantitative variables. Be sure to include the following:   |  |                         |    |                                |     |          |    |           |    |
| a. What level of measurement is required for each variable type?   |  |                         |    |                                |     |          |    |           |    |
| b. Can both types be used to describe both samples and populations?  |  |                         |    |                                |     |          |    |           |    |
| 24. Describe the similarities and differences between a frequency table and a frequency distribution. Be sure to include which requires qualitative data and which requires quantitative data.   |  |                         |    |                                |     |          |    |           |    |
| 25. Alexandra Damonte will be building a new resort in Myrtle Beach, South Carolina. She must decide how to design the resort based on the type of activities that the resort will offer to its customers. A recent poll of 300 potential customers showed the following results about customers' preferences for planned resort activities: |  |                         |    |                                |     |          |    |           |    |
|  | <table border="1"><tbody><tr><td>Like planned activities</td><td>63</td></tr><tr><td>Do not like planned activities</td><td>135</td></tr><tr><td>Not sure</td><td>78</td></tr><tr><td>No answer</td><td>24</td></tr></tbody></table> | Like planned activities | 63 | Do not like planned activities | 135 | Not sure | 78 | No answer | 24 |
| Like planned activities  | 63   |                         |    |                                |     |          |    |           |    |
| Do not like planned activities   | 135  |                         |    |                                |     |          |    |           |    |
| Not sure   | 78   |                         |    |                                |     |          |    |           |    |
| No answer  | 24   |                         |    |                                |     |          |    |           |    |
| a. What is the table called?   |  |                         |    |                                |     |          |    |           |    |
| b. Draw a bar chart to portray the survey results.   |  |                         |    |                                |     |          |    |           |    |

### Data Analytics

The goal of the Data Analytics sections is to develop analytical skills. The exercises present a real-world context with supporting data. The data sets are printed in Appendix A and available to download from Connect. Statistical software is required to analyze the data and respond to the exercises. Each data set is used to explore questions and discover findings that relate to a real-world context. For each business context, a story is uncovered as students progress from chapter 1 to 15.

| DATA ANALYTICS   |
|--|
| (The data for these exercises are available in Connect.)   |
| 51. <b>FILE</b> Refer to the North Valley Real Estate data, which report information on homes sold during the last year. For the variable <i>price</i> , select an appropriate class interval and organize the selling prices into a frequency distribution. Write a brief report summarizing your findings. Be sure to answer the following questions in your report. |
| a. Around what values of price do the data tend to cluster?  |
| b. Based on the frequency distribution, what is the typical selling price in the first class? What is the typical selling price in the last class?   |
| c. Draw a cumulative relative frequency distribution. Using this distribution, fifty percent of the homes sold for what price or less? Estimate the lower price of the top ten percent of homes sold. About what percent of the homes sold for less than \$300,000?  |

## Practice Test

The Practice Test is intended to give students an idea of content that might appear on a test and how the test might be structured. The Practice Test includes both objective questions and problems covering the material studied in the section.

**PRACTICE TEST**

**Part 1—Objective**

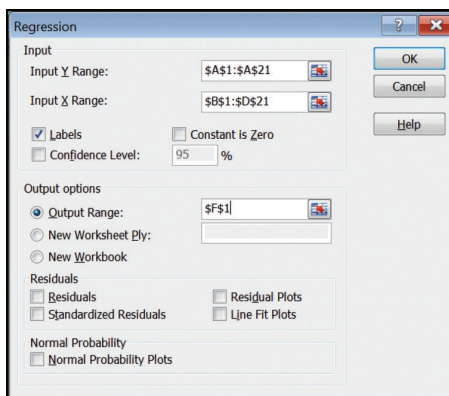
1. A grouping of *qualitative data* into mutually exclusive classes showing the number of observations in each class is known as a \_\_\_\_\_.
2. A grouping of *quantitative data* into mutually exclusive classes showing the number of observations in each class is known as a \_\_\_\_\_.
3. A graph in which the classes for qualitative data are reported on the horizontal axis and the class frequencies (proportional to the heights of the bars) on the vertical axis is called a \_\_\_\_\_.
4. A circular chart that shows the proportion or percentage that each class represents of the total is called a \_\_\_\_\_.
5. A graph in which the classes of a quantitative variable are marked on the horizontal axis and the class frequencies on the vertical axis is called a \_\_\_\_\_.
6. A set of data included 70 observations. How many classes would you suggest to construct a frequency distribution?  
\_\_\_\_\_

## APPENDIX MATERIAL

### Software Commands

Software examples using Excel, MegaStat, and Minitab are included throughout the text. The explanations of the computer input commands are placed at the end of the text in Appendix C.

- 14–1.** The Excel commands to produce the multiple regression output on page 422 are:
- a. Import the data from Connect. The file name is **Tbl14**.
  - b. Select the **Data** tab on the top menu. Then on the far right, select **Data Analysis**. Select **Regression** and click **OK**.
  - c. Make the **Input Y Range** *A1:A21*, the **Input X Range** *B1:D21*, check the **Labels** box, the **Output Range** is *F1*, then click **OK**.



## Answers to Self-Review

The worked-out solutions to the Self-Reviews are provided at the end of the text in Appendix E.

- 2–3**
- a.  $2^6 = 64 < 73 < 128 = 2^7$ , so seven classes are recommended.
  - b. The interval width should be at least  $(488 - 320)/7 = 24$ . Class intervals of either 25 or 30 are reasonable.
  - c. Assuming a class interval of 25 and beginning with a lower limit of 300, eight classes are required. If we use an interval of 30 and begin with a lower limit of 300, only seven classes are required. Seven classes is the better alternative.

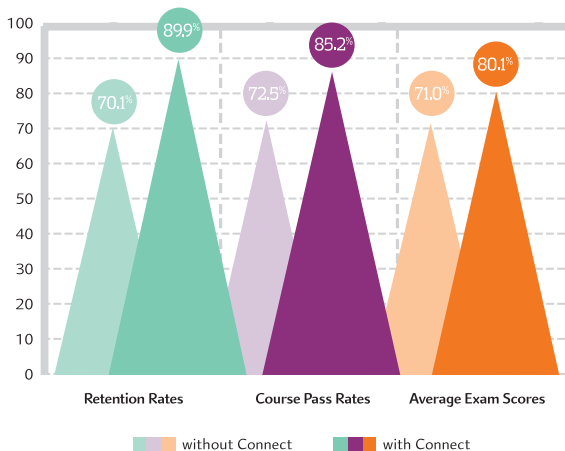
| Distance Classes | Frequency | Percent |
|------------------|-----------|---------|
| 300 up to 330    | 2         | 2.7%    |
| 330 up to 360    | 2         | 2.7     |
| 360 up to 390    | 17        | 23.3    |
| 390 up to 420    | 27        | 37.0    |
| 420 up to 450    | 22        | 30.1    |
| 450 up to 480    | 1         | 1.4     |
| 480 up to 510    | 2         | 2.7     |
| Grand Total      | 73        | 100.00  |

McGraw-Hill Connect® is a highly reliable, easy-to-use homework and learning management solution that utilizes learning science and award-winning adaptive tools to improve student results.

## Homework and Adaptive Learning

- Connect's assignments help students contextualize what they've learned through application, so they can better understand the material and think critically.
- Connect will create a personalized study path customized to individual student needs through SmartBook®.
- SmartBook helps students study more efficiently by delivering an interactive reading experience through adaptive highlighting and review.

Connect's Impact on Retention Rates, Pass Rates, and Average Exam Scores



Over **7 billion** questions have been answered, making McGraw-Hill Education products more intelligent, reliable, and precise.

Using **Connect** improves retention rates by **19.8%**, passing rates by **12.7%**, and exam scores by **9.1%**.

73% of instructors who use **Connect** require it; instructor satisfaction increases by 28% when **Connect** is required.

## Quality Content and Learning Resources

- Connect content is authored by the world's best subject matter experts, and is available to your class through a simple and intuitive interface.
- The Connect eBook makes it easy for students to access their reading material on smartphones and tablets. They can study on the go and don't need internet access to use the eBook as a reference, with full functionality.
- Multimedia content such as videos, simulations, and games drive student engagement and critical thinking skills.



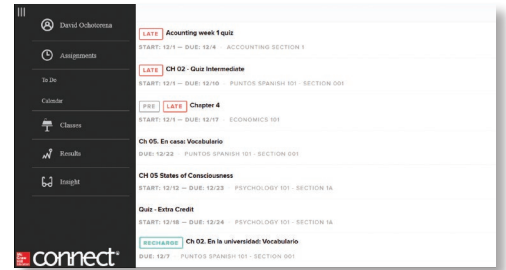
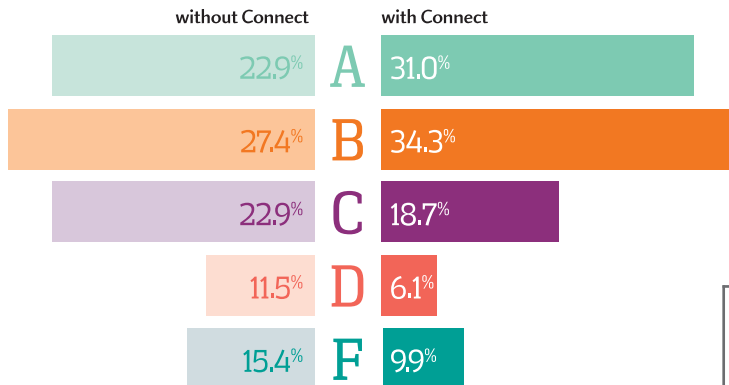
# Robust Analytics and Reporting

- Connect Insight® generates easy-to-read reports on individual students, the class as a whole, and on specific assignments.
- The Connect Insight dashboard delivers data on performance, study behavior, and effort. Instructors can quickly identify students who struggle and focus on material that the class has yet to master.
- Connect automatically grades assignments and quizzes, providing easy-to-read reports on individual and class performance.



©Hero Images/Getty Images

## Impact on Final Course Grade Distribution



More students earn  
**As and Bs** when they  
use **Connect**.

## Trusted Service and Support

- Connect integrates with your LMS to provide single sign-on and automatic syncing of grades. Integration with Blackboard®, D2L®, and Canvas also provides automatic syncing of the course calendar and assignment-level linking.
- Connect offers comprehensive service, support, and training throughout every phase of your implementation.
- If you're looking for some guidance on how to use Connect, or want to learn tips and tricks from super users, you can find tutorials as you work. Our Digital Faculty Consultants and Student Ambassadors offer insight into how to achieve the results you want with Connect.

## INSTRUCTOR LIBRARY

The McGraw-Hill Education Connect Business Statistics Instructor Library is your repository for additional resources to improve student engagement in and out of class. You can select and use any asset that enhances your lecture, including:

- **Solutions Manual** The Solutions Manual, carefully revised by the authors, contains solutions to all basic, intermediate, and challenge problems found at the end of each chapter.
- **Test Bank** The Test Bank, revised by Wendy Bailey of Troy University, contains hundreds of true/false, multiple choice, and short-answer/discussions, updated based on the revisions of the authors. The level of difficulty varies, as indicated by the easy, medium, and difficult labels.
- **PowerPoint Presentations** Prepared by Stephanie Campbell of Mineral Area College, the presentations contain exhibits, tables, key points, and summaries in a visually stimulating collection of slides.
- **Excel Templates** There are templates for various end-of-chapter problems that have been set as Excel spreadsheets—all denoted by an icon. Students can easily download and save the files and use the data to solve end-of-chapter problems.

## MEGASTAT<sup>®</sup> FOR MICROSOFT EXCEL<sup>®</sup>

MegaStat by J. B. Orris of Butler University is a full-featured Excel statistical analysis add-in that is available on the MegaStat website at [www.mhhe.com/megastat](http://www.mhhe.com/megastat) (for purchase). MegaStat works with recent versions of Microsoft Excel (Windows and Mac OS X). See the website for details on supported versions.

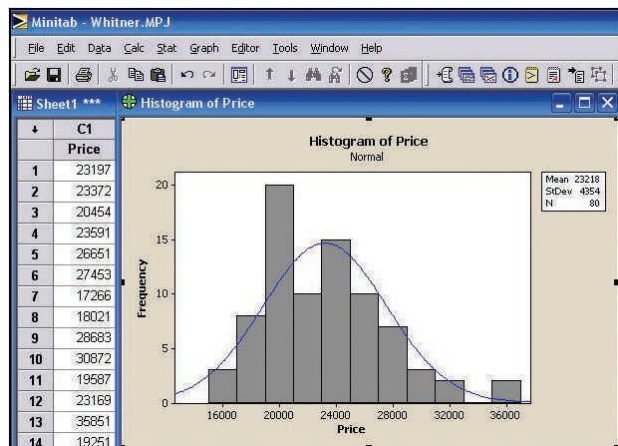
Once installed, MegaStat will always be available on the Excel add-ins ribbon with no expiration date or data limitations. MegaStat performs statistical analyses within an Excel workbook. When a MegaStat menu item is selected, a dialog box pops up for data selection and options. Since MegaStat is an easy-to-use extension of Excel, students can focus on learning statistics without being distracted by the software. Ease-of-use features include Auto Expand for quick data selection and Auto Label detect.

MegaStat does most calculations found in introductory statistics textbooks, such as computing descriptive statistics, creating frequency distributions, and computing probabilities, as well as hypothesis testing, ANOVA, chi-square analysis, and regression analysis (simple and multiple). MegaStat output is carefully formatted and appended to an output worksheet.

Video tutorials are included that provide a walk-through using MegaStat for typical business statistics topics. A context-sensitive help system is built into MegaStat, and a User's Guide is included in PDF format.

## MINITAB<sup>®</sup>/SPSS<sup>®</sup>/JMP<sup>®</sup>

Minitab Version 17, SPSS Student Version 18.0, and JMP Student Edition Version 8 are software products that are available to help students solve the exercises with data files. Each software product can be packaged with any McGraw-Hill business statistics text.



# ACKNOWLEDGMENTS

This edition of *Basic Statistics for Business and Economics* is the product of many people: students, colleagues, reviewers, and the staff at McGraw-Hill Education. We thank them all. We wish to express our sincere gratitude to the reviewers:

Stefan Ruediger

*Arizona State University*

Anthony Clark

*St. Louis Community College*

Umair Khalil

*West Virginia University*

Leonie Stone

*SUNY Geneseo*

Golnaz Taghvatalab

*Central Michigan University*

John Yarber

*Northeast Mississippi Community  
College*

John Beyers

*University of Maryland*

Mohammad Kazemi

*University of North Carolina  
Charlotte*

Anna Terzyan

*Loyola Marymount University*

Lee O. Cannell

*El Paso Community College*

Their suggestions and thorough reviews of the previous edition and the manuscript of this edition make this a better text.

Special thanks go to a number of people. Shelly Moore, College of Western Idaho, and John Arcaro, Lakeland Community College, accuracy checked the Connect exercises. Ed Pappanastos, Troy University, built new data sets and revised SmartBook. Rene Ordonez, Southern Oregon University, built the Connect guided examples. Wendy Bailey, Troy University, prepared the test bank. Stephanie Campbell, Mineral Area College, prepared the PowerPoint decks. Vickie Fry, Westmoreland County Community College, provided countless hours of digital accuracy checking and support.

We also wish to thank the staff at McGraw-Hill Education. This includes Noelle Bathurst, Portfolio Manager; Michele Janicek, Lead Product Developer; Ryan McAndrews, Product Developer; Lori Koettters, Content Project Manager; Daryl Horrocks, Program Manager; and others we do not know personally, but who have made valuable contributions.



# ENHANCEMENTS TO BASIC STATISTICS FOR BUSINESS & ECONOMICS, 9E

## CHANGES MADE TO INDIVIDUAL CHAPTERS

### CHAPTER 1 What Is Statistics?

- Revised Self-Review 1–2.
- New section describing business analytics and its integration with the text.
- Updated Exercises 2, 3, 17, and 19.
- New photo and chapter opening exercise.
- New introduction with new graphic showing the increasing amount of information collected and processed with new technologies.
- New ordinal scale example based on rankings of states by business climate.
- The chapter includes several new examples.
- Chapter is more focused on the revised learning objectives, improving the chapter's flow.
- Revised Exercise 17 is based on economic data.
- New Data Analytics section with new data and questions.

### CHAPTER 2 Describing Data: Frequency Tables, Frequency Distributions, and Graphic Presentation

- Revised chapter introduction.
- Added more explanation about cumulative relative frequency distributions.
- Updated Exercises 38, 45, 47, and 48 using real data.
- Revised Self-Review 2–3 to include data.
- New Data Analytics section with new data and questions.

### CHAPTER 3 Describing Data: Numerical Measures

- Updated Self-Review 3–2.
- Reorganized chapter based on revised learning objectives.
- Replaced the mean deviation with more emphasis on the variance and standard deviation.
- Updated Statistics in Action.
- New Data Analytics section with new data and questions.

### CHAPTER 4 Describing Data: Displaying and Exploring Data

- Updated Exercise 22 with 2016 New York Yankee player salaries.
- New Data Analytics section with new data and questions.

### CHAPTER 5 A Survey of Probability Concepts

- Updated Exercises 39 and 52 using real data.
- New explanation of odds compared to probabilities.
- New Exercise 21.
- New Example/Solution for demonstrating contingency tables and tree diagrams.

- New contingency table in Exercise 31.
- Revised Example/Solution demonstrating the combination formula.
- New Data Analytics section with new data and questions.

### CHAPTER 6 Discrete Probability Distributions

- Expanded discussion of random variables.
- Revised the Example/Solution in the section on Poisson distribution.
- Updated Exercise 18 and added new Exercises 54, 55, and 56.
- Revised the section on the binomial distribution.
- Revised Example/Solution demonstrating the binomial distribution.
- New exercise using a raffle at a local golf club to demonstrate probability and expected returns.
- New Data Analytics section with new data and questions.

### CHAPTER 7 Continuous Probability Distributions

- Revised Self-Review 7–1.
- Revised the Example/Solutions using Uber as the context.
- Updated Exercises 19, 22, 28, 37, and 48.
- Updated Statistics in Action.
- Revised Self-Review 7–2 based on daily personal water consumption.
- Revised explanation of the Empirical Rule as it relates to the normal distribution.
- New Data Analytics section with new data and questions.

### CHAPTER 8 Sampling Methods and the Central Limit Theorem

- New example of simple random sampling and the application of the table of random numbers.
- The discussions of systematic random, stratified random, and cluster sampling have been revised.
- Revised Exercise 44 based on the price of a gallon of milk.
- New Data Analytics section with new data and questions.

### CHAPTER 9 Estimation and Confidence Intervals

- New Self-Review 9–3 problem description.
- Updated Exercises 5, 6, 12, 14, 24, 37, 39, and 55.
- New Statistics in Action describing EPA fuel economy.
- New separate section on point estimates.
- Integration and application of the central limit theorem.
- A revised simulation demonstrating the interpretation of confidence level.
- New presentation on using the  $t$  table to find  $z$  values.
- A revised discussion of determining the confidence interval for the population mean.

- Expanded section on calculating sample size.
- New Data Analytics section with new data and questions.

### CHAPTER 10 One-Sample Tests of Hypothesis

- Revised the Example/Solutions using an airport cell phone parking lot as the context.
- Revised software solution and explanation of  $p$ -values.
- Conducting a test of hypothesis about a population proportion is moved to Chapter 15.
- New example introducing the concept of hypothesis testing.
- Sixth step added to the hypothesis testing procedure emphasizing the interpretation of the hypothesis test results.
- New Data Analytics section with new data and questions.

### CHAPTER 11 Two-Sample Tests of Hypothesis

- Updated Exercises 5, 9, 30, and 44.
- New introduction to the chapter.
- Section of two-sample tests about proportions moved to Chapter 15.
- Changed subscripts in Example/Solution for easier understanding.
- New Data Analytics section with new data and questions.

### CHAPTER 12 Analysis of Variance

- Revised Self-Reviews 12–1 and 12–3.
- Updated Exercises 10, 16, 25, and 30.
- New introduction to the chapter.
- New Exercise 16 using the speed of browsers to search the Internet.
- Revised Exercise 25 comparing learning in traditional versus online courses.
- New section on comparing two population variances.
- New example illustrating the comparison of variances.
- Revised the names of the airlines in the one-way ANOVA example.
- Changed the subscripts in Example/Solution for easier understanding.
- New Data Analytics section with new data and questions.

### CHAPTER 13 Correlation and Linear Regression

- Added new conceptual formula to relate the standard error to the regression ANOVA table.
- Updated Exercises 41 and 57.
- Rewrote the introduction section to the chapter.
- The data used as the basis for the North American Copier Sales Example/Solution used throughout the chapter have been changed and expanded to 15 observations to more clearly demonstrate the chapter's learning objectives.
- Revised section on transforming data using the economic relationship between price and sales.
- New Exercises 35 (transforming data), 36 (Masters prizes and scores), 43 (2012 NFL points scored versus points allowed), 44 (store size and sales), and 61 (airline distance and fare).
- New Data Analytics section with new data and questions.

### CHAPTER 14 Multiple Regression Analysis

- Updated Exercises 19, 22, and 25.
- Rewrote the section on evaluating the multiple regression equation.
- More emphasis on the regression ANOVA table.
- Enhanced the discussion of the  $p$ -value in decision making.
- More emphasis on calculating the variance inflation factor to evaluate multicollinearity.
- New Data Analytics section with new data and questions.

### CHAPTER 15 Nonparametric Methods: Nominal-Level Hypothesis Tests

- Updated the context of Manelli Perfume Company Example/Solution.
- Revised the "Hypothesis Test of Unequal Expected Frequencies" Example/Solution.
- Moved one-sample and two-sample tests of proportions from Chapters 10 and 11 to Chapter 15.
- New example introducing goodness-of-fit tests.
- Removed the graphical methods to evaluate normality.
- Revised section on contingency table analysis with a new Example/Solution.
- New Data Analytics section with new data and questions.



# BRIEF CONTENTS

- 1 What is Statistics? 1
  - 2 Describing Data: Frequency Tables, Frequency Distributions, and Graphic Presentation 19
  - 3 Describing Data: Numerical Measures 53
  - 4 Describing Data: Displaying and Exploring Data 88
  - 5 A Survey of Probability Concepts 117
  - 6 Discrete Probability Distributions 155
  - 7 Continuous Probability Distributions 184
  - 8 Sampling Methods and the Central Limit Theorem 210
  - 9 Estimation and Confidence Intervals 242
  - 10 One-Sample Tests of Hypothesis 274
  - 11 Two-Sample Tests of Hypothesis 305
  - 12 Analysis of Variance 334
  - 13 Correlation and Linear Regression 365
  - 14 Multiple Regression Analysis 418
  - 15 Nonparametric Methods:  
Nominal-Level Hypothesis Tests 469
- Appendixes:**
- Data Sets, Tables, Software Commands, Answers 503
  - Glossary 578
  - Index 581



*A Note from the Authors* vi

*Preface* vii

## 1 What is Statistics? 1

Introduction 2

Why Study Statistics? 2

What is Meant by Statistics? 3

Types of Statistics 4

Descriptive Statistics 4

Inferential Statistics 5

Types of Variables 6

Levels of Measurement 7

Nominal-Level Data 7

Ordinal-Level Data 8

Interval-Level Data 9

Ratio-Level Data 10

**EXERCISES** 11

Ethics and Statistics 12

Basic Business Analytics 12

Chapter Summary 13

Chapter Exercises 14

Data Analytics 17

Practice Test 17

## 2 Describing Data: FREQUENCY TABLES, FREQUENCY DISTRIBUTIONS, AND GRAPHIC PRESENTATION 19

Introduction 20

Constructing Frequency Tables 20

Relative Class Frequencies 21

Graphic Presentation  
of Qualitative Data 22

**EXERCISES** 26

Constructing Frequency Distributions 27

Relative Frequency Distribution 31

**EXERCISES** 32

Graphic Presentation of a Distribution 33

Histogram 33

Frequency Polygon 36

**EXERCISES** 38

Cumulative Distributions 39

**EXERCISES** 42

Chapter Summary 43

Chapter Exercises 44

Data Analytics 51

Practice Test 51

## 3 Describing Data: NUMERICAL MEASURES 53

Introduction 54

Measures of Location 54

The Population Mean 55

The Sample Mean 56

Properties of the Arithmetic Mean 57

**EXERCISES** 58

The Median 59

The Mode 61

**EXERCISES** 63

The Relative Positions of the Mean,  
Median, and Mode 64

**EXERCISES** 65

Software Solution 66

The Weighted Mean 67

**EXERCISES** 68

Why Study Dispersion? 68

Range 69

Variance 70

**EXERCISES** 72

Population Variance 73

Population Standard Deviation 75

**EXERCISES** 75

Sample Variance and Standard  
Deviation 76

Software Solution 77

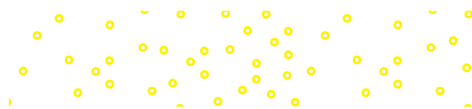
**EXERCISES** 78

Interpretation and Uses of the Standard  
Deviation 78

Chebyshev's Theorem 78

The Empirical Rule 79

**EXERCISES** 80



|   |     |   |     |
|---|-----|---|-----|
| Ethics and Reporting Results                      | 81  | Rules of Multiplication to Calculate Probability                                  | 132 |
| Chapter Summary                                   | 81  | Special Rule of Multiplication  | 132 |
| Pronunciation Key                                 | 82  | General Rule of Multiplication  | 133 |
| Chapter Exercises                                 | 83  | Contingency Tables  | 135 |
| Data Analytics                                    | 86  | Tree Diagrams   | 138 |
| Practice Test                                     | 86  | <b>EXERCISES</b>  | 140 |
| <b>4 Describing Data:</b>                         |     | Principles of Counting  | 142 |
| <b>DISPLAYING AND EXPLORING DATA</b>              | 88  | The Multiplication Formula  | 142 |
| Introduction                                      | 89  | The Permutation Formula   | 143 |
| Dot Plots   | 89  | The Combination Formula   | 145 |
| <b>EXERCISES</b>                                  | 91  | <b>EXERCISES</b>  | 147 |
| Measures of Position                              | 92  | Chapter Summary   | 147 |
| Quartiles, Deciles, and Percentiles               | 92  | Pronunciation Key   | 148 |
| <b>EXERCISES</b>                                  | 96  | Chapter Exercises   | 148 |
| Box Plots   | 96  | Data Analytics  | 153 |
| <b>EXERCISES</b>                                  | 99  | Practice Test   | 154 |
| Skewness  | 100 | <b>6 Discrete Probability</b>   |     |
| <b>EXERCISES</b>                                  | 103 | <b>Distributions</b>  | 155 |
| Describing the Relationship between Two Variables | 104 | Introduction  | 156 |
| Contingency Tables                                | 106 | What is a Probability Distribution?   | 156 |
| <b>EXERCISES</b>                                  | 108 | Random Variables  | 158 |
| Chapter Summary                                   | 109 | Discrete Random Variable  | 159 |
| Pronunciation Key                                 | 110 | Continuous Random Variable  | 160 |
| Chapter Exercises                                 | 110 | The Mean, Variance, and Standard Deviation of a Discrete Probability Distribution | 160 |
| Data Analytics                                    | 115 | Mean  | 160 |
| Practice Test                                     | 115 | Variance and Standard Deviation   | 160 |
|   |     | <b>EXERCISES</b>  | 162 |
|   |     | Binomial Probability Distribution   | 164 |
| <b>5 A Survey of Probability Concepts</b>         | 117 | How is a Binomial Probability Computed?   | 165 |
| Introduction                                      | 118 | Binomial Probability Tables   | 167 |
| What is a Probability?                            | 119 | <b>EXERCISES</b>  | 170 |
| Approaches to Assigning Probabilities             | 121 | Cumulative Binomial Probability Distributions                                     | 171 |
| Classical Probability                             | 121 | <b>EXERCISES</b>  | 172 |
| Empirical Probability                             | 122 | Poisson Probability Distribution  | 173 |
| Subjective Probability                            | 124 | <b>EXERCISES</b>  | 178 |
| <b>EXERCISES</b>                                  | 125 | Chapter Summary   | 178 |
| Rules of Addition for Computing Probabilities     | 126 | Chapter Exercises   | 179 |
| Special Rule of Addition                          | 126 | Data Analytics  | 183 |
| Complement Rule                                   | 128 | Practice Test   | 183 |
| The General Rule of Addition                      | 129 |   |     |
| <b>EXERCISES</b>                                  | 131 |   |     |

## 7 Continuous Probability Distributions 184

Introduction 185

The Family of Uniform Probability Distributions 185

**EXERCISES** 188

The Family of Normal Probability Distributions 189

The Standard Normal Probability Distribution 192

Applications of the Standard Normal Distribution 193

The Empirical Rule 193

**EXERCISES** 195

Finding Areas under the Normal Curve 196

**EXERCISES** 199

**EXERCISES** 201

**EXERCISES** 204

Chapter Summary 204

Chapter Exercises 205

Data Analytics 208

Practice Test 209

## 8 Sampling Methods and the Central Limit Theorem 210

Introduction 211

Sampling Methods 211

Reasons to Sample 211

Simple Random Sampling 212

Systematic Random Sampling 215

Stratified Random Sampling 215

Cluster Sampling 216

**EXERCISES** 217

Sampling “Error” 219

Sampling Distribution of the Sample Mean 221

**EXERCISES** 224

The Central Limit Theorem 225

**EXERCISES** 231

Using the Sampling Distribution of the Sample Mean 232

**EXERCISES** 234

Chapter Summary 235

Pronunciation Key 236

Chapter Exercises 236

Data Analytics 241

Practice Test 241

## 9 Estimation and Confidence Intervals 242

Introduction 243

Point Estimate for a Population Mean 243

Confidence Intervals for a Population Mean 244

Population Standard Deviation, Known  $\sigma$  244

A Computer Simulation 249

**EXERCISES** 251

Population Standard Deviation,  $\sigma$  Unknown 252

**EXERCISES** 259

A Confidence Interval for a Population

Proportion 260

**EXERCISES** 263

Choosing an Appropriate Sample Size 263

Sample Size to Estimate a Population Mean 264

Sample Size to Estimate a Population

Proportion 265

**EXERCISES** 267

Chapter Summary 267

Chapter Exercises 268

Data Analytics 272

Practice Test 273

## 10 One-Sample Tests of Hypothesis 274

Introduction 275

What is Hypothesis Testing? 275

Six-Step Procedure for Testing a Hypothesis 276

Step 1: State the Null Hypothesis ( $H_0$ ) and the Alternate Hypothesis ( $H_1$ ) 276

Step 2: Select a Level of Significance 277

Step 3: Select the Test Statistic 279

Step 4: Formulate the Decision Rule 279

Step 5: Make a Decision 280

Step 6: Interpret the Result 280

One-Tailed and Two-Tailed Hypothesis Tests 281

Hypothesis Testing for a Population Mean: Known Population Standard Deviation 283

A Two-Tailed Test 283

A One-Tailed Test 286

$p$ -Value in Hypothesis Testing 287

**EXERCISES** 289

Hypothesis Testing for a Population Mean: Population Standard Deviation Unknown 290

**EXERCISES** 295

A Statistical Software Solution 296

**EXERCISES** 297

|                   |     |
|-------------------|-----|
| Chapter Summary   | 299 |
| Pronunciation Key | 299 |
| Chapter Exercises | 300 |
| Data Analytics    | 303 |
| Practice Test     | 303 |

## 11 Two-Sample Tests of Hypothesis 305

|  |     |
|--|-----|
| Introduction   | 306 |
| Two-Sample Tests of Hypothesis: Independent Samples                    | 306 |
| <b>EXERCISES</b>   | 311 |
| Comparing Population Means with Unknown Population Standard Deviations | 312 |
| Two-Sample Pooled Test   | 312 |
| <b>EXERCISES</b>   | 316 |
| Two-Sample Tests of Hypothesis: Dependent Samples                      | 318 |
| Comparing Dependent and Independent Samples                            | 321 |
| <b>EXERCISES</b>   | 324 |
| Chapter Summary  | 325 |
| Pronunciation Key  | 326 |
| Chapter Exercises  | 326 |
| Data Analytics   | 332 |
| Practice Test  | 332 |

## 12 Analysis of Variance 334

|  |     |
|--|-----|
| Introduction                                       | 335 |
| Comparing Two Population Variances                 | 335 |
| The $F$ Distribution                               | 335 |
| Testing a Hypothesis of Equal Population Variances | 336 |
| <b>EXERCISES</b>                                   | 339 |
| ANOVA: Analysis of Variance                        | 340 |
| ANOVA Assumptions                                  | 340 |
| The ANOVA Test                                     | 342 |
| <b>EXERCISES</b>                                   | 349 |
| Inferences about Pairs of Treatment Means          | 350 |
| <b>EXERCISES</b>                                   | 352 |
| Chapter Summary                                    | 354 |
| Pronunciation Key                                  | 355 |
| Chapter Exercises                                  | 355 |
| Data Analytics                                     | 362 |
| Practice Test                                      | 363 |

## 13 Correlation and Linear Regression 365

|   |     |
|---|-----|
| Introduction  | 366 |
| What is Correlation Analysis?   | 366 |
| The Correlation Coefficient   | 369 |
| <b>EXERCISES</b>  | 374 |
| Testing the Significance of the Correlation Coefficient   | 376 |
| <b>EXERCISES</b>  | 379 |
| Regression Analysis   | 380 |
| Least Squares Principle   | 380 |
| Drawing the Regression Line   | 383 |
| <b>EXERCISES</b>  | 386 |
| Testing the Significance of the Slope   | 388 |
| <b>EXERCISES</b>  | 390 |
| Evaluating a Regression Equation's Ability to Predict   | 391 |
| The Standard Error of Estimate  | 391 |
| The Coefficient of Determination  | 392 |
| <b>EXERCISES</b>  | 393 |
| Relationships among the Correlation Coefficient, the Coefficient of Determination, and the Standard Error of Estimate | 393 |
| <b>EXERCISES</b>  | 395 |
| Interval Estimates of Prediction  | 396 |
| Assumptions Underlying Linear Regression  | 396 |
| Constructing Confidence and Prediction Intervals  | 397 |
| <b>EXERCISES</b>  | 400 |
| Transforming Data   | 400 |
| <b>EXERCISES</b>  | 403 |
| Chapter Summary   | 404 |
| Pronunciation Key   | 406 |
| Chapter Exercises   | 406 |
| Data Analytics  | 415 |
| Practice Test   | 416 |

## 14 Multiple Regression Analysis 418

|   |     |
|---|-----|
| Introduction                              | 419 |
| Multiple Regression Analysis              | 419 |
| <b>EXERCISES</b>                          | 423 |
| Evaluating a Multiple Regression Equation | 425 |
| The ANOVA Table                           | 425 |



Multiple Standard Error of Estimate 426  
 Coefficient of Multiple Determination 427  
 Adjusted Coefficient of Determination 428

**EXERCISES** 429

**Inferences in Multiple Linear Regression** 429

Global Test: Testing the Multiple  
 Regression Model 429  
 Evaluating Individual Regression Coefficients 432

**EXERCISES** 435

**Evaluating the Assumptions of Multiple  
 Regression** 436

Linear Relationship 437  
 Variation in Residuals Same for Large  
 and Small  $\hat{y}$  Values 438  
 Distribution of Residuals 439  
 Multicollinearity 439  
 Independent Observations 441

**Qualitative Independent Variables** 442

**Stepwise Regression** 445

**EXERCISES** 447

**Review of Multiple Regression** 448

**Chapter Summary** 454

**Pronunciation Key** 455

**Chapter Exercises** 456

**Data Analytics** 466

**Practice Test** 467

**Goodness-of-Fit Tests: Comparing Observed and  
 Expected Frequency Distributions** 479

Hypothesis Test of Equal Expected  
 Frequencies 479

**EXERCISES** 484

Hypothesis Test of Unequal Expected  
 Frequencies 486

**Limitations of Chi-Square** 487

**EXERCISES** 489

**Contingency Table Analysis** 490

**EXERCISES** 493

**Chapter Summary** 494

**Pronunciation Key** 495

**Chapter Exercises** 495

**Data Analytics** 500

**Practice Test** 501

**APPENDIXES** 503

*Appendix A: Data Sets* 504

*Appendix B: Tables* 513

*Appendix C: Software Commands* 526

*Appendix D: Answers to Odd-Numbered  
 Chapter Exercises* 534

*Solutions to Practice Tests* 566

*Appendix E: Answers to Self-Review* 570

*Glossary* 578

*Index* 581

*Key Formulas*

*Student's t Distribution*

*Areas under the Normal Curve*

## 15 Nonparametric Methods:

### NOMINAL-LEVEL HYPOTHESIS

**TESTS** 469

**Introduction** 470

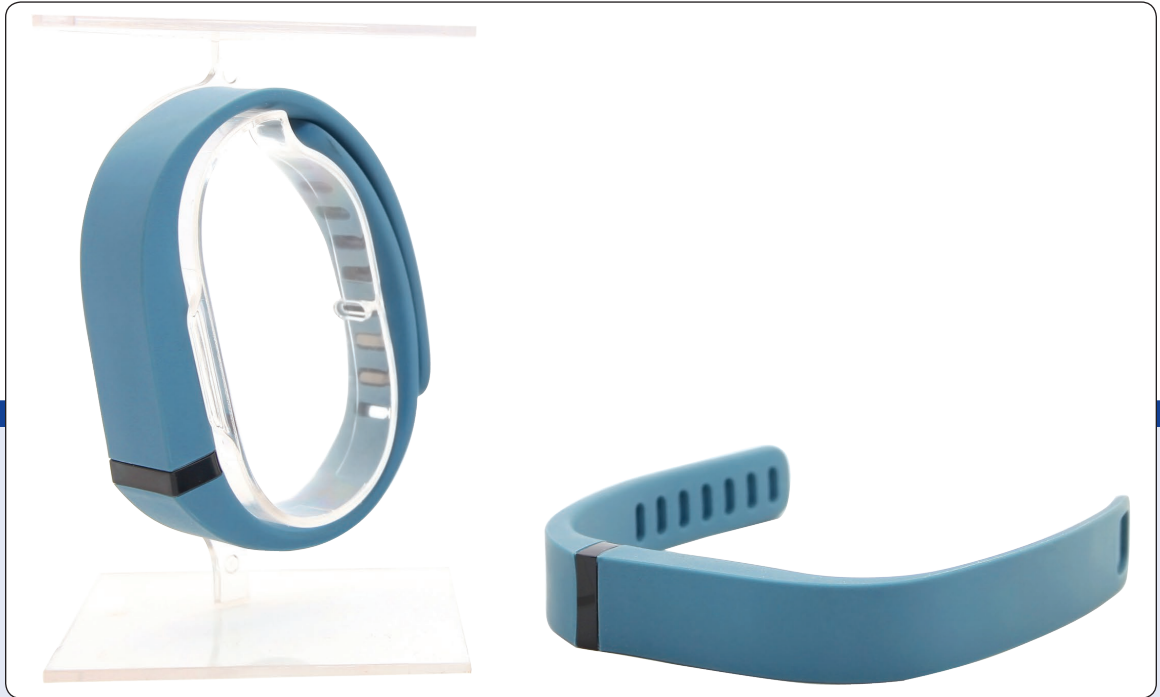
**Test a Hypothesis of a Population  
 Proportion** 470

**EXERCISES** 473

**Two-Sample Tests about Proportions** 474

**EXERCISES** 478

# What is Statistics?



©Kelvin Wong/Shutterstock

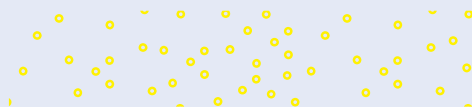
- ▲ **BEST BUY** sells Fitbit wearable technology products that track a person's physical activity and sleep quality. The Fitbit technology collects daily information on the number of steps per day so a person can track calories consumed. The information can be synced with a cell phone and displayed with a Fitbit app. Assume you know the daily number of Fitbit Flex 2 units sold last month at the Best Buy store in Collegeville, Pennsylvania. Describe a situation where the number of units sold is considered a sample. Illustrate a second situation where the number of units sold is considered a population. (See Exercise 11 and **LO1-3**.)

## LEARNING OBJECTIVES

---

*When you have completed this chapter, you will be able to:*

- LO1-1** Explain why knowledge of statistics is important.
- LO1-2** Define statistics and provide an example of how statistics is applied.
- LO1-3** Differentiate between descriptive and inferential statistics.
- LO1-4** Classify variables as qualitative or quantitative, and discrete or continuous.
- LO1-5** Distinguish between nominal, ordinal, interval, and ratio levels of measurement.
- LO1-6** List the values associated with the practice of statistics.





©Gregor Schuster/Getty Images RF

## INTRODUCTION

Suppose you work for a large company and your supervisor asks you to decide if a new version of a smartphone should be produced and sold. You start by thinking about the product's innovations and new features. Then, you stop and realize the consequences of the decision. The product will need to make a profit, so the pricing and the costs of production and distribution are all very important. The decision to introduce the product is based on many alternatives. So how will you know? Where do you start?

Without extensive experience in the industry, beginning to develop an intelligence that will make you an expert is essential. You select three other people to work with and meet with them. The conversation focuses on what you need to know and what information and data you need. In your meeting, many questions are asked. How many competitors are already in the market? How are smartphones priced? What design features do competitors' products have? What features does the market require? What do customers want in a smartphone? What do customers like about the existing products? The answers will be based on business intelligence consisting of data and information collected through customer surveys, engineering analysis, and market research. In the end, your presentation to support your decision regarding the introduction of a new smartphone is based on the statistics that you use to summarize and organize your data, the statistics that you use to compare the new product to existing products, and the statistics to estimate future sales, costs, and revenues. The statistics will be the focus of the conversation that you will have with your supervisor about this very important decision.

As a decision maker, you will need to acquire and analyze data to support your decisions. The purpose of this text is to develop your knowledge of basic statistical techniques and methods and how to apply them to develop the business and personal intelligence that will help you make decisions.

### LO1-1

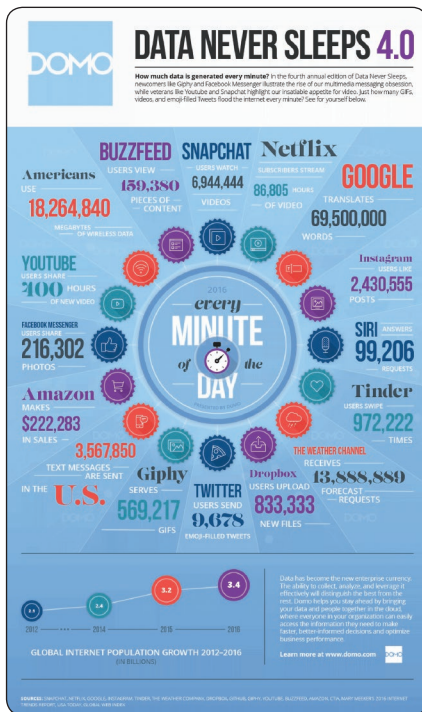
Explain why knowledge of statistics is important.

## WHY STUDY STATISTICS?

If you look through your university catalogue, you will find that statistics is required for many college programs. As you investigate a future career in accounting, economics, human resources, finance, business analytics, or other business area, you also will discover that statistics is required as part of these college programs. So why is statistics a requirement in so many disciplines?

A major driver of the requirement for statistics knowledge is the technologies available for capturing data. Examples include the technology that Google uses to track how Internet users access websites. As people use Google to search the Internet, Google records every search and then uses these data to sort and prioritize the results for future Internet searches. One recent estimate indicates that Google processes 20,000 terabytes of information per day. Big-box retailers like Target, Walmart, Kroger, and others scan every purchase and use the data to manage the distribution of products, to make decisions about marketing and sales, and to track daily and even hourly sales. Police departments collect and use data to provide city residents with maps that communicate information about crimes committed and their location. Every organization is collecting and using data to develop knowledge and intelligence that will help people make informed decisions, and to track the implementation of their decisions. The graphic to the left shows the amount of data generated every minute ([www.domo.com](http://www.domo.com)). A good working knowledge of statistics is useful for summarizing and organizing data to provide information that is useful and supportive of decision making. Statistics is used to make valid comparisons and to predict the outcomes of decisions.

In summary, there are at least three reasons for studying statistics: (1) data are collected everywhere and require statistical knowledge to



Courtesy of Domo.com, Josh James, "Data Never Sleeps 4.0," June 28, 2016

make the information useful, (2) statistical techniques are used to make professional and personal decisions, and (3) no matter what your career, you will need a knowledge of statistics to understand the world and to be conversant in your career. An understanding of statistics and statistical method will help you make more effective personal and professional decisions.

**LO1-2**

Define statistics and provide an example of how statistics is applied.

**STATISTICS IN ACTION**

A feature of our textbook is called *Statistics in Action*. Read each one carefully to get an appreciation of the wide application of statistics in management, economics, nursing, law enforcement, sports, and other disciplines.

- In 2017, *Forbes* published a list of the richest Americans. William Gates, founder of Microsoft Corporation, is the richest. His net worth is estimated at \$86.0 billion. ([www.forbes.com](http://www.forbes.com))
- In 2017, the four largest privately owned American companies, ranked by revenue, were Cargill, Koch Industries, State Farm Mutual Automobile Insurance, and Dell. ([www.forbes.com](http://www.forbes.com))
- In the United States, a typical high school graduate earns \$668 per week, a typical college graduate with a bachelor's degree earns \$1,101 per week, and a typical college graduate with a master's degree earns \$1,326 per week. ([www.bls.gov/emp/ep\\_chart\\_001.htm](http://www.bls.gov/emp/ep_chart_001.htm))

**WHAT IS MEANT BY STATISTICS?**

This question can be rephrased in two, subtly different ways: what are statistics and what is statistics? To answer the first question, a statistic is a number used to communicate a piece of information. Examples of **statistics** are:

- The inflation rate is 2%.
- Your grade point average is 3.5.
- The price of a new Tesla Model S sedan is \$79,570.

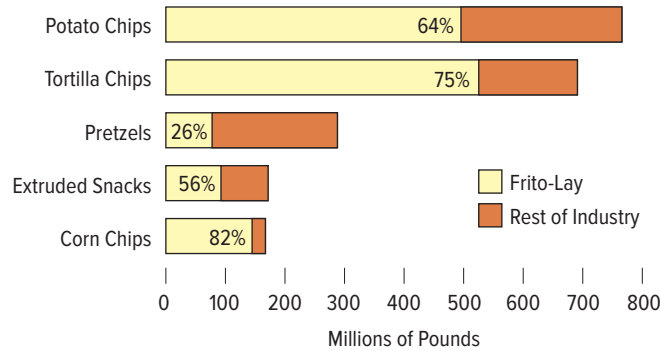
Each of these statistics is a numerical fact and communicates a very limited piece of information that is not very useful by itself. However, if we recognize that each of these statistics is part of a larger discussion, then the question “what **is** statistics” is applicable. Statistics is the set of knowledge and skills used to organize, summarize, and analyze data. The results of statistical analysis will start interesting conversations in the search for knowledge and intelligence that will help us make decisions. For example:

- The inflation rate for the calendar year was 0.7%. By applying statistics we could compare this year's inflation rate to the past observations of inflation. Is it higher, lower, or about the same? Is there a trend of increasing or decreasing inflation? Is there a relationship between interest rates and government bonds?
- Your grade point average (GPA) is 3.5. By collecting data and applying statistics, you can determine the required GPA to be admitted to the Master of Business Administration program at the University of Chicago, Harvard University, or the University of Michigan. You can determine the likelihood that you would be admitted to a particular program. You may be interested in interviewing for a management position with Procter & Gamble. What GPA does Procter & Gamble require for college graduates with a bachelor's degree? Is there a range of acceptable GPAs?
- You are budgeting for a new car. You would like to own an electric car with a small carbon footprint. The price of the Tesla Model S sedan is \$79,570. By collecting additional data and applying statistics, you can analyze the alternatives. For example, another choice is a hybrid car that runs on both gas and electricity such as a 2017 Toyota Prius. It can be purchased for about \$28,659. Another hybrid, the Chevrolet Volt, costs \$33,995. What are the differences in the cars' specifications? What additional information can be collected and summarized so that you can make a good purchase decision?

Another example of using statistics to provide information to evaluate decisions is the distribution and market share of Frito-Lay products. Data are collected on each of the Frito-Lay product lines. These data include the market share and the pounds of product sold. Statistics is used to present this information in a bar chart in Chart 1–1. It clearly shows Frito-Lay's dominance in the potato, corn, and tortilla chip markets. It also shows the absolute measure of pounds of each product line consumed in the United States.

These examples show that statistics is more than the presentation of numerical information. Statistics is about collecting and processing information to create a conversation, to stimulate additional questions, and to provide a basis for making decisions. Specifically, we define **statistics** as:

**STATISTICS** The science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions.



**CHART 1-1** Frito-Lay Volume and Share of Major Snack Chip Categories in U.S. Supermarkets

In this book, you will learn the basic techniques and applications of statistics that you can use to support your decisions, both personal and professional. To start, we will differentiate between descriptive and inferential statistics.

### LO1-3

Differentiate between descriptive and inferential statistics.

## TYPES OF STATISTICS

When we use statistics to generate information for decision making from data, we use either descriptive statistics or inferential statistics. Their application depends on the questions asked and the type of data available.

### Descriptive Statistics

Masses of unorganized data—such as the census of population, the weekly earnings of thousands of computer programmers, and the individual responses of 2,000 registered voters regarding their choice for president of the United States—are of little value as is. However, descriptive statistics can be used to organize data into a meaningful form. We define **descriptive statistics** as:

**DESCRIPTIVE STATISTICS** Methods of organizing, summarizing, and presenting data in an informative way.

The following are examples that apply descriptive statistics to summarize a large amount of data and provide information that is easy to understand.

- There are a total of 46,837 miles of interstate highways in the United States. The interstate system represents only 1% of the nation's total roads but carries more than 20% of the traffic. The longest is I-90, which stretches from Boston to Seattle, a distance of 3,099 miles. The shortest is I-878 in New York City, which is 0.70 mile in length. Alaska does not have any interstate highways, Texas has the most interstate miles at 3,232, and New York has the most interstate routes with 28.
- The average person spent \$147 on traditional Valentine's Day merchandise in 2016. This is an increase of \$5 from 2015. As is typical of most years, men spent about twice as much as women. Men typically spent an average of \$196, whereas women spent an average of \$100 ([www.fundivo.com](http://www.fundivo.com)).

Statistical methods and techniques to generate descriptive statistics are presented in Chapters 2 and 4. These include organizing and summarizing data with frequency distributions and presenting frequency distributions with charts and graphs. In addition, statistical measures to summarize the characteristics of a distribution are discussed in Chapter 3.

## Inferential Statistics

Sometimes we must make decisions based on a limited set of data. For example, we would like to know the operating characteristics, such as fuel efficiency measured by miles per gallon, of sport utility vehicles (SUVs) currently in use. If we spent a lot of time, money, and effort, all the owners of SUVs could be surveyed. In this case, our goal would be to survey the **population** of SUV owners.

**POPULATION** The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest.

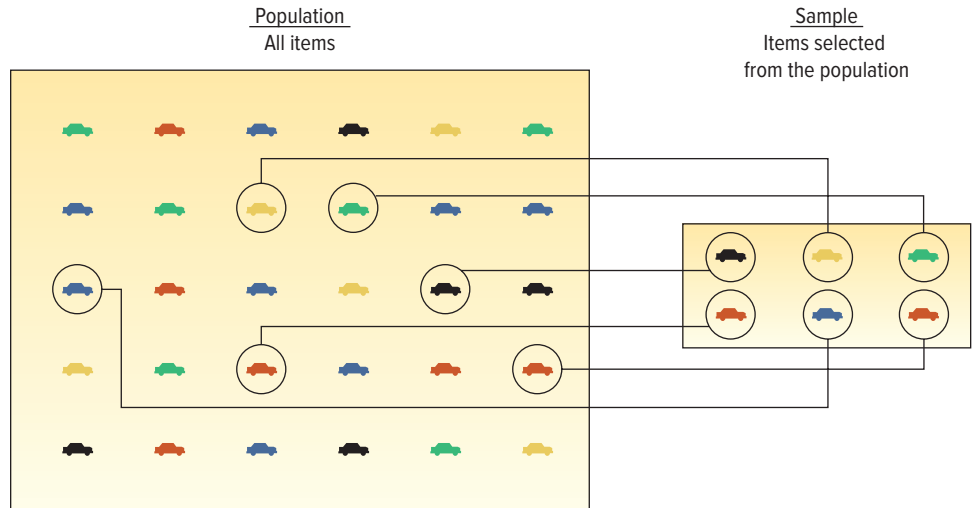
However, based on inferential statistics, we can survey a limited number of SUV owners and collect a **sample** from the population.

**SAMPLE** A portion, or part, of the population of interest.

Samples often are used to obtain reliable estimates of population parameters. (Sampling is discussed in Chapter 8.) In the process, we make trade-offs between the time, money, and effort to collect the data and the error of estimating a population parameter. The process of sampling SUVs is illustrated in the following graphic. In this example, we would like to know the mean or average SUV fuel efficiency. To estimate the mean of the population, six SUVs are sampled and the mean of their MPG is calculated.

### STATISTICS IN ACTION

Where did statistics get its start? In 1662 John Graunt published an article called “Natural and Political Observations Made upon Bills of Mortality.” The author’s “observations” were the result of a study and analysis of a weekly church publication called “Bill of Mortality,” which listed births, christenings, and deaths and their causes. Graunt realized that the Bills of Mortality represented only a fraction of all births and deaths in London. However, he used the data to reach broad conclusions or inferences about the impact of disease, such as the plague, on the general population. His logic is an example of statistical inference. His analysis and interpretation of the data are thought to mark the start of statistics.



So, the sample of six SUVs represents evidence from the population that we use to reach an inference or conclusion about the average MPG for all SUVs. The process of sampling from a population with the objective of estimating properties of a population is called **inferential statistics**.

**INFERENCEAL STATISTICS** The methods used to estimate a property of a population on the basis of a sample.

Inferential statistics is widely applied to learn something about a population in business, agriculture, politics, and government, as shown in the following examples:

- Television networks constantly monitor the popularity of their programs by hiring Nielsen and other organizations to sample the preferences of TV viewers. For example, *NCIS* was the most watched show during the week of March 13–19, 2017. A total of 14.16 million viewers watched this show ([www.nielsen.com](http://www.nielsen.com)). These program ratings are used to make decisions about advertising rates and whether to continue or cancel a program.
- In 2017, a sample of U.S. Internal Revenue Service tax preparation volunteers were tested with three standard tax returns. The sample indicated that tax returns were completed with a 49% accuracy rate. In other words, there were errors on about half of the returns. In this example, the statistics are used to make decisions about how to improve the accuracy rate by correcting the most common errors and improving the training of volunteers.

*A feature of our text is self-review problems. There are a number of them interspersed throughout each chapter. The first self-review follows. Each self-review tests your comprehension of preceding material. The answer and method of solution are given in Appendix E. You can find the answer to the following self-review in 1–1 in Appendix E. We recommend that you solve each one and then check your answer.*

## SELF-REVIEW 1–1



The answers are in Appendix E.

The Atlanta-based advertising firm Brandon and Associates asked a sample of 1,960 consumers to try a newly developed chicken dinner by Boston Market. Of the 1,960 sampled, 1,176 said they would purchase the dinner if it is marketed.

- Is this an example of descriptive statistics or inferential statistics? Explain.
- What could Brandon and Associates report to Boston Market regarding acceptance of the chicken dinner in the population?

### LO1-4

Classify variables as qualitative or quantitative, and discrete or continuous.

## TYPES OF VARIABLES

There are two basic types of variables: (1) qualitative and (2) quantitative (see Chart 1–2). When an object or individual is observed and recorded as a nonnumeric characteristic, it is a qualitative variable or an attribute. Examples of qualitative variables are gender, beverage preference, type of vehicle owned, state of birth, and eye color. When a variable is quantitative, we usually count the number of observations for each category and determine

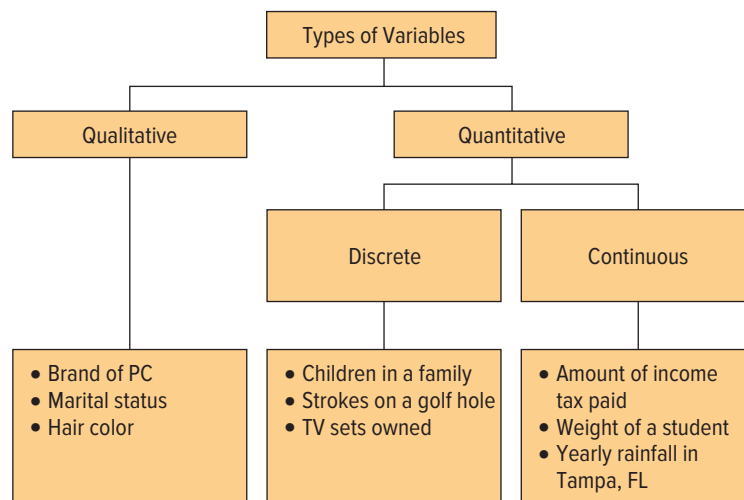


CHART 1–2 Summary of the Types of Variables

what percent are in each category. For example, if we observe the variable eye color, what percent of the population has blue eyes and what percent has brown eyes? If the variable is type of vehicle, what percent of the total number of cars sold last month were SUVs? Qualitative variables are often summarized in charts and bar graphs (Chapter 2).

When a variable can be reported numerically, it is called a quantitative variable. Examples of quantitative variables are the balance in your checking account, the number of gigabytes of data used on your cell phone plan last month, the life of a car battery (such as 42 months), and the number of people employed by a company.

Quantitative variables are either discrete or continuous. Discrete variables can assume only certain values, and there are “gaps” between the values. Examples of discrete variables are the number of bedrooms in a house (1, 2, 3, 4, etc.), the number of cars arriving at Exit 25 on I-4 in Florida near Walt Disney World in an hour (326, 421, etc.), and the number of students in each section of a statistics course (25 in section A, 42 in section B, and 18 in section C). We count, for example, the number of cars arriving at Exit 25 on I-4, and we count the number of statistics students in each section. Notice that a home can have 3 or 4 bedrooms, but it cannot have 3.56 bedrooms. Thus, there is a “gap” between possible values. Typically, discrete variables are counted.

Observations of a continuous variable can assume any value within a specific range. Examples of continuous variables are the air pressure in a tire and the weight of a shipment of tomatoes. Other examples are the ounces of raisins in a box of raisin bran cereal and the duration of flights from Orlando to San Diego. Grade point average (GPA) is a continuous variable. We could report the GPA of a particular student as 3.2576952. The usual practice is to round to 3 places—3.258. Typically, continuous variables result from measuring.

#### LO1-5

Distinguish between nominal, ordinal, interval, and ratio levels of measurement.



©McGraw-Hill Education

## LEVELS OF MEASUREMENT

Data can be classified according to levels of measurement. The level of measurement determines how data should be summarized and presented. It also will indicate the type of statistical analysis that can be performed. Here are two examples of the relationship between measurement and how we apply statistics. There are six colors of candies in a bag of M&Ms. Suppose we assign brown a value of 1, yellow 2, blue 3, orange 4, green 5, and red 6. What kind of variable is the color of an M&M? It is a qualitative variable.

Suppose someone summarizes M&M color by adding the assigned color values, divides the sum by the number of M&Ms, and reports that the mean color is 3.56. How do we interpret this statistic? You are correct in concluding that it has no meaning as a measure of M&M color. As a qualitative variable, we can only report the count and percentage of each color in a bag of M&Ms. As a second example, in a high school track meet there are eight competitors in the 400-meter run. We report the order of finish and that the mean finish is 4.5. What does the mean finish tell us? Nothing! In both of these instances, we have not used the appropriate statistics for the level of measurement.

There are four levels of measurement: nominal, ordinal, interval, and ratio. The lowest, or the most primitive, measurement is the nominal level. The highest is the ratio level of measurement.

### Nominal-Level Data

For the **nominal level of measurement**, observations of a qualitative variable are measured and recorded as labels or names. The labels or names can only be classified and counted. There is no particular order to the labels.

**NOMINAL LEVEL OF MEASUREMENT** Data recorded at the nominal level of measurement are represented as labels or names. They have no order. They can only be classified and counted.



The classification of the six colors of M&M milk chocolate candies is an example of the nominal level of measurement. We simply classify the candies by color. There is no natural order. That is, we could report the brown candies first, the orange first, or any of the other colors first. Recording the variable gender is another example of the nominal level of measurement. Suppose we count the number of students entering a football game with a student ID and report how many are men and how many are women. We could report either the men or the women first. For the data measured at the nominal level, we are limited to counting the number in each category of the variable. Often, we convert these counts to percentages. For example, a random sample of M&M candies reports the following percentages for each color:

| Color  | Percent in a bag |
|--------|------------------|
| Blue   | 24%              |
| Green  | 20%              |
| Orange | 16%              |
| Yellow | 14%              |
| Red    | 13%              |
| Brown  | 13%              |

To process the data for a variable measured at the nominal level, we often numerically code the labels or names. For example, if we are interested in measuring the home state for students at East Carolina University, we would assign a student's home state of Alabama a code of 1, Alaska a code of 2, Arizona a 3, and so on. Using this procedure with an alphabetical listing of states, Wisconsin is coded 49 and Wyoming 50. Realize that the number assigned to each state is still a label or name. The reason we assign numerical codes is to facilitate counting the number of students from each state with statistical software. Note that assigning numbers to the states does not give us license to manipulate the codes as numerical information. Specifically, in this example,  $1 + 2 = 3$  corresponds to Alabama + Alaska = Arizona. Clearly, the nominal level of measurement does not permit any mathematical operation that has any valid interpretation.

## Ordinal-Level Data

The next higher level of measurement is the **ordinal level**. For this level of measurement, a qualitative variable or attribute is either ranked or rated on a relative scale.

**ORDINAL LEVEL OF MEASUREMENT** Data recorded at the ordinal level of measurement are based on a relative ranking or rating of items based on a defined attribute or qualitative variable. Variables based on this level of measurement are only ranked or counted.

### Best Business Climate

1. Florida
2. Utah
3. Texas
4. Georgia
5. Indiana
6. Tennessee
7. Nebraska
8. North Carolina
9. Virginia
10. Washington

For example, many businesses make decisions about where to locate their facilities; in other words, where is the best place for their business? Business Facilities ([www.businessfacilities.com](http://www.businessfacilities.com)) publishes a list of the top 10 states for the "best business climate." The 2016 rankings are shown to the left. They are based on the evaluation of many different factors, including the cost of labor, business tax climate, quality of life, transportation infrastructure, educated workforce, and economic growth potential.

This is an example of an ordinal scale because the states are ranked in order of best to worst business climate. That is, we know the relative order of the states based

on the attribute. For example, in 2016 Florida had the best business climate and Utah was second. Indiana was fifth, and that was better than Tennessee but not as good as Georgia. Notice we cannot say that Florida's business climate is five times better than Indiana's business climate because the magnitude of the differences between the states is not known. To put it another way, we do not know if the magnitude of the difference between Louisiana and Utah is the same as between Texas and Georgia.

Another example of the ordinal-level measure is based on a scale that measures an attribute. This type of scale is used when students rate instructors on a variety of attributes. One attribute may be: "Overall, how do you rate the quality of instruction in this class?" A student's response is recorded on a relative scale of inferior, poor, good, excellent, and superior. An important characteristic of using a relative measurement scale is that we cannot distinguish the magnitude of the differences between groups. We do not know if the difference between "Superior" and "Good" is the same as the difference between "Poor" and "Inferior."

Table 1–1 lists the frequencies of 60 student ratings of instructional quality for Professor James Brunner in an Introduction to Finance course. The data are summarized based on the order of the scale used to rate the instructor. That is, they are summarized by the number of students who indicated a rating of superior (6), good (26), and so on. We also can convert the frequencies to percentages. About 43.3% (26/60) of the students rated the instructor as good.

**TABLE 1–1** Rating of a Finance Professor

| Rating   | Frequency | Percentage |
|----------|-----------|------------|
| Superior | 6         | 10.0%      |
| Good     | 26        | 43.3%      |
| Average  | 16        | 26.7%      |
| Poor     | 9         | 15.0%      |
| Inferior | 3         | 5.0%       |

## Interval-Level Data

The **interval level of measurement** is the next highest level. It includes all the characteristics of the ordinal level, but, in addition, the difference or interval between values is meaningful.

**INTERVAL LEVEL OF MEASUREMENT** For data recorded at the interval level of measurement, the interval or the distance between values is meaningful. The interval level of measurement is based on a scale with a known unit of measurement.

The Fahrenheit temperature scale is an example of the interval level of measurement. Suppose the high temperatures on three consecutive winter days in Boston are 28, 31, and 20 degrees Fahrenheit. These temperatures can be easily ranked, but we can also determine the interval or distance between temperatures. This is possible because 1 degree Fahrenheit represents a constant unit of measurement. That is, the distance between 10 and 15 degrees Fahrenheit is 5 degrees, and is the same as the 5-degree distance between 50 and 55 degrees Fahrenheit. It is also important to note that 0 is just a point on the scale. It does not represent the absence of the condition. The measurement of zero degrees Fahrenheit does not represent the absence of heat or cold. But by our own measurement scale, it is cold! A major limitation of a variable measured at the interval level is that we cannot make statements similar to 20 degrees Fahrenheit is twice as warm as 10 degrees Fahrenheit.

Another example of the interval scale of measurement is women's dress sizes. Listed below is information on several dimensions of a standard U.S. woman's dress.

| Size | Bust (in) | Waist (in) | Hips (in) |
|------|-----------|------------|-----------|
| 8    | 32        | 24         | 35        |
| 10   | 34        | 26         | 37        |
| 12   | 36        | 28         | 39        |
| 14   | 38        | 30         | 41        |
| 16   | 40        | 32         | 43        |
| 18   | 42        | 34         | 45        |
| 20   | 44        | 36         | 47        |
| 22   | 46        | 38         | 49        |
| 24   | 48        | 40         | 51        |
| 26   | 50        | 42         | 53        |
| 28   | 52        | 44         | 55        |

Why is the "size" scale an interval measurement? Observe that as the size changes by two units (say from size 10 to size 12 or from size 24 to size 26), each of the measurements increases by 2 inches. To put it another way, the intervals are the same.

There is no natural zero point for dress size. A "size 0" dress does not have "zero" material. Instead, it would have a 24-inch bust, 16-inch waist, and 27-inch hips. Moreover, the ratios are not reasonable. If you divide a size 28 by a size 14, you do not get the same answer as dividing a size 20 by a size 10. Neither ratio is equal to two, as the "size" number would suggest. In short, if the distances between the numbers make sense, but the ratios do not, then you have an interval scale of measurement.

## Ratio-Level Data

Almost all quantitative variables are recorded on the **ratio level of measurement**. The ratio level is the "highest" level of measurement. It has all the characteristics of the interval level, but, in addition, the 0 point and the ratio between two numbers are both meaningful.

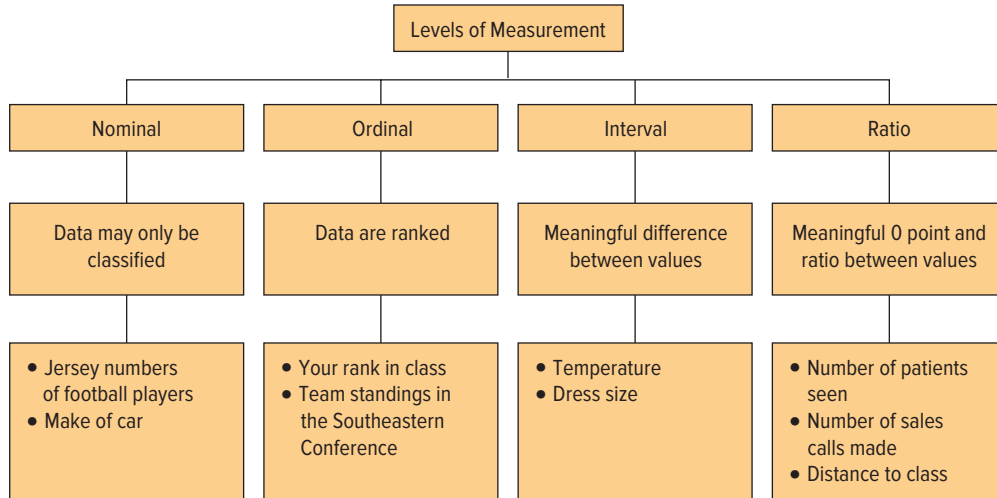
**RATIO LEVEL OF MEASUREMENT** Data recorded at the ratio level of measurement are based on a scale with a known unit of measurement and a meaningful interpretation of zero on the scale.

Examples of the ratio scale of measurement include wages, units of production, weight, changes in stock prices, distance between branch offices, and height. Money is also a good illustration. If you have zero dollars, then you have no money, and a wage of \$50 per hour is two times the wage of \$25 per hour. Weight also is measured at the ratio level of measurement. If a scale is correctly calibrated, then it will read 0 when nothing is on the scale. Further, something that weighs 1 pound is half as heavy as something that weighs 2 pounds.

Table 1–2 illustrates the ratio scale of measurement for the variable, annual income for four father-and-son combinations. Observe that the senior Lahey earns twice as much as his son. In the Rho family, the son makes twice as much as the father.

**TABLE 1–2** Father–Son Income Combinations

| Name   | Father   | Son       |
|--------|----------|-----------|
| Lahey  | \$80,000 | \$ 40,000 |
| Nale   | 90,000   | 30,000    |
| Rho    | 60,000   | 120,000   |
| Steele | 75,000   | 130,000   |



**CHART 1–3** Summary and Examples of the Characteristics for Levels of Measurement

Chart 1–3 summarizes the major characteristics of the various levels of measurement. The level of measurement will determine the type of statistical methods that can be used to analyze a variable. Statistical methods to analyze variables measured on a nominal level are discussed in Chapter 15. Statistical methods to analyze variables measured on an interval or ratio level are presented in Chapters 9 through 14.

## SELF-REVIEW 1–2



- (a) The mean age of people who listen to talk radio is 42.1 years. What level of measurement is used to assess the variable age?
- (b) In a survey of luxury-car owners, 8% of the U.S. population owned luxury cars. In California and Georgia, 14% of people owned luxury cars. Two variables are included in this information. What are they and how are they measured?

## EXERCISES

The answers to the odd-numbered exercises are in Appendix D.

- What is the level of measurement for each of the following variables?
  - Student IQ ratings.
  - Distance students travel to class.
  - The jersey numbers of a sorority soccer team.
  - A student's state of birth.
  - A student's academic class—that is, freshman, sophomore, junior, or senior.
  - Number of hours students study per week.
- Slate* is a daily magazine on the Web. Its business activities can be described by a number of variables. What is the level of measurement for each of the following variables?
  - The number of hits on their website on Saturday between 8:00 a.m. and 9:00 a.m.
  - The departments, such as food and drink, politics, foreign policy, sports, etc.
  - The number of weekly hits on the Sam's Club ad.
  - The number of years each employee has been employed with *Slate*.
- On the Web, go to your favorite news source and find examples of each type of variable. Write a brief memo that lists the variables and describes them in terms of qualitative or quantitative, discrete or continuous, and the measurement level.

4. For each of the following, determine whether the group is a sample or a population.
  - a. The participants in a study of a new cholesterol drug.
  - b. The drivers who received a speeding ticket in Kansas City last month.
  - c. People on welfare in Cook County (Chicago), Illinois.
  - d. The 30 stocks that make up the Dow Jones Industrial Average.

**LO1-6**

List the values associated with the practice of statistics.

## ETHICS AND STATISTICS

Following events such as Wall Street money manager Bernie Madoff's Ponzi scheme, which swindled billions from investors, and financial misrepresentations by Enron and Tyco, business students need to understand that these events were based on the misrepresentation of business and financial information. In each case, people within each organization reported financial information to investors that indicated the companies were performing much better than the actual situation. When the true financial information was reported, the companies were worth much less than advertised. The result was many investors lost all or nearly all of the money they had invested.

The article "Statistics and Ethics: Some Advice for Young Statisticians," in *The American Statistician* 57, no. 1 (2003), offers guidance. The authors advise us to practice statistics with integrity and honesty, and urge us to "do the right thing" when collecting, organizing, summarizing, analyzing, and interpreting numerical information. The real contribution of statistics to society is a moral one. Financial analysts need to provide information that truly reflects a company's performance so as not to mislead individual investors. Information regarding product defects that may be harmful to people must be analyzed and reported with integrity and honesty. The authors of *The American Statistician* article further indicate that when we practice statistics, we need to maintain "an independent and principled point-of-view" when analyzing and reporting findings and results.

As you progress through this text, we will highlight ethical issues in the collection, analysis, presentation, and interpretation of statistical information. We also hope that, as you learn about using statistics, you will become a more informed consumer of information. For example, you will question a report based on data that do not fairly represent the population, a report that does not include all relevant statistics, one that includes an incorrect choice of statistical measures, and a presentation that introduces bias in an attempt to mislead or misrepresent.

## BASIC BUSINESS ANALYTICS

A knowledge of statistics is necessary to support the increasing need for companies and organizations to apply business analytics. Business analytics is used to process and analyze data and information to support a story or narrative of a company's business, such as "what makes us profitable," or "how will our customers respond to a change in marketing?" In addition to statistics, an ability to use computer software to summarize, organize, analyze, and present the findings of statistical analysis is essential. In this text, we will be using very elementary applications of business analytics using common and available computer software. Throughout our text, we will use Microsoft Excel and, occasionally, Minitab. Universities and colleges usually offer access to Microsoft Excel. Your computer already may be packaged with Microsoft Excel. If not, the Microsoft Office package with Excel often is sold at a reduced academic price through your university or college. In this text, we use Excel for the majority of the applications. We also use an Excel "add-in" called MegaStat. If your instructor requires this package, it is available at [www.mhhe.com/megastat](http://www.mhhe.com/megastat). This add-in gives Excel the capability to produce additional statistical reports. Occasionally, we use Minitab to illustrate an application. See [www.minitab.com](http://www.minitab.com) for further information. Minitab also offers discounted academic pricing. The 2016 version of Microsoft Excel supports the analyses in our text. However,

earlier versions of Excel for Apple Mac computers do not have the necessary add-in. If you do not have Excel 2016 and are using an Apple Mac computer with Excel, you can download the free, trial version of StatPlus at [www.analystsoft.com](http://www.analystsoft.com). It is a statistical software package that will integrate with Excel for Mac computers.

The following example shows the application of Excel to perform a statistical summary. It refers to sales information from the Applewood Auto Group, a multi-location car sales and service company. The Applewood information has sales information for 180 vehicle sales. Each sale is described by several variables: the age of the buyer, whether the buyer is a repeat customer, the location of the dealership for the sale, the type of vehicle sold, and the profit for the sale. The following shows Excel's summary of statistics for the variable profit. The summary of profit shows the mean profit per vehicle was \$1,843.17, the median profit was slightly more at \$1,882.50, and profit ranged from \$294 to \$3,292.

| APPLEWOOD AUTO GROUP.xlsx |     |         |           |              |          |                    |               |           |
|---------------------------|-----|---------|-----------|--------------|----------|--------------------|---------------|-----------|
|                           | A   | B       | C         | D            | E        | F                  | G             | H         |
| 1                         | Age | Profit  | Location  | Vehicle-Type | Previous |                    | <i>Profit</i> |           |
| 2                         | 33  | \$1,889 | Olean     | SUV          | 1        |                    |               |           |
| 3                         | 47  | \$1,461 | Kane      | Sedan        | 0        | Mean               |               | 1843.17   |
| 4                         | 44  | \$1,532 | Tionesta  | SUV          | 3        | Standard Error     |               | 47.97     |
| 5                         | 53  | \$1,220 | Olean     | Sedan        | 0        | Median             |               | 1882.50   |
| 6                         | 51  | \$1,674 | Sheffield | Sedan        | 1        | Mode               |               | 1915.00   |
| 7                         | 41  | \$2,389 | Kane      | Truck        | 1        | Standard Deviation |               | 643.63    |
| 8                         | 58  | \$2,058 | Kane      | SUV          | 1        | Sample Variance    |               | 414256.61 |
| 9                         | 35  | \$1,919 | Tionesta  | SUV          | 1        | Kurtosis           |               | -0.22     |
| 10                        | 45  | \$1,266 | Olean     | Sedan        | 0        | Skewness           |               | -0.24     |
| 11                        | 54  | \$2,991 | Tionesta  | Sedan        | 0        | Range              |               | 2998      |
| 12                        | 56  | \$2,695 | Kane      | Sedan        | 2        | Minimum            |               | 294       |
| 13                        | 41  | \$2,165 | Tionesta  | SUV          | 0        | Maximum            |               | 3292      |
| 14                        | 38  | \$1,766 | Sheffield | SUV          | 0        | Sum                |               | 331770    |
| 15                        | 48  | \$1,952 | Tionesta  | Compact      | 1        | Count              |               | 180       |

Source: Microsoft Excel

Throughout the text, we will encourage the use of computer software to summarize, describe, and present information and data. The applications of Excel are supported by instructions so that you can learn how to apply Excel to do statistical analysis. The instructions are presented in Appendix C of this text. These data and other data sets and files are available in Connect.

## CHAPTER SUMMARY

- I. Statistics is the science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions.
- II. There are two types of statistics.
  - A. Descriptive statistics are procedures used to organize and summarize data.
  - B. Inferential statistics involve taking a sample from a population and making estimates about a population based on the sample results.
    1. A population is an entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest.
    2. A sample is a part of the population.
- III. There are two types of variables.
  - A. A qualitative variable is nonnumeric.
    1. Usually we are interested in the number or percent of the observations in each category.
    2. Qualitative data are usually summarized in graphs and bar charts.

- B. There are two types of quantitative variables and they are usually reported numerically.
  - 1. Discrete variables can assume only certain values, and there are usually gaps between values.
  - 2. A continuous variable can assume any value within a specified range.
- IV. There are four levels of measurement.
  - A. With the nominal level, the data are sorted into categories with no particular order to the categories.
  - B. The ordinal level of measurement presumes that one classification is ranked higher than another.
  - C. The interval level of measurement has the ranking characteristic of the ordinal level of measurement plus the characteristic that the distance between values is a constant size.
  - D. The ratio level of measurement has all the characteristics of the interval level, plus there is a 0 point and the ratio of two values is meaningful.

## CHAPTER EXERCISES

- 5. Explain the difference between qualitative and quantitative variables. Give an example of qualitative and quantitative variables.
- 6. Explain the difference between a sample and a population.
- 7. Explain the difference between a discrete and a continuous variable. Give an example of each not included in the text.
- 8. For the following situations, would you collect information using a sample or a population? Why?
  - a. Statistics 201 is a course taught at a university. Professor Rauch has taught nearly 1,500 students in the course over the past 5 years. You would like to know the average grade for the course.
  - b. As part of a research project, you need to report the average profit as a percentage of revenue for the #1-ranked corporation in the Fortune 500 for each of the last 10 years.
  - c. You are looking forward to graduation and your first job as a salesperson for one of five large pharmaceutical corporations. Planning for your interviews, you will need to know about each company's mission, profitability, products, and markets.
  - d. You are shopping for a new MP3 music player such as the Apple iPod. The manufacturers advertise the number of music tracks that can be stored in the memory. Usually, the advertisers assume relatively short, popular songs to estimate the number of tracks that can be stored. You, however, like Broadway musical tunes and they are much longer. You would like to estimate how many Broadway tunes will fit on your MP3 player.
- 9. Exits along interstate highways were formerly numbered successively from the western or southern border of a state. However, the Department of Transportation has recently changed most of them to agree with the numbers on the mile markers along the highway.
  - a. What level of measurement were data on the consecutive exit numbers?
  - b. What level of measurement are data on the milepost numbers?
  - c. Discuss the advantages of the newer system.
- 10. A poll solicits a large number of college undergraduates for information on the following variables: the name of their cell phone provider (AT&T, Verizon, and so on), the numbers of minutes used last month (200, 400, for example), and their satisfaction with the service (Terrible, Adequate, Excellent, and so forth). What is the level of measurement for each of these three variables?
- 11. Best Buy sells Fitbit wearable technology products that track a person's activity. For example, the Fitbit technology collects daily information on the number of steps per day so a person can track calories consumed. The information can be synced with a cell phone and displayed with a Fitbit app. Assume you know the daily number of Fitbit Flex

2 units sold last month at the Best Buy store in Collegeville, Pennsylvania. Describe a situation where the number of units sold is considered a sample. Illustrate a second situation where the number of units sold is considered a population.

12. Using the concepts of sample and population, describe how a presidential election is unlike an “exit” poll of the electorate.
13. Place these variables in the following classification tables. For each table, summarize your observations and evaluate if the results are generally true. For example, salary is reported as a continuous quantitative variable. It is also a continuous ratio-scaled variable.
  - a. Salary
  - b. Gender
  - c. Sales volume of MP3 players
  - d. Soft drink preference
  - e. Temperature
  - f. SAT scores
  - g. Student rank in class
  - h. Rating of a finance professor
  - i. Number of home video screens

|              | Discrete Variable | Continuous Variable |
|--------------|-------------------|---------------------|
| Qualitative  |                   |                     |
| Quantitative |                   | a. Salary           |

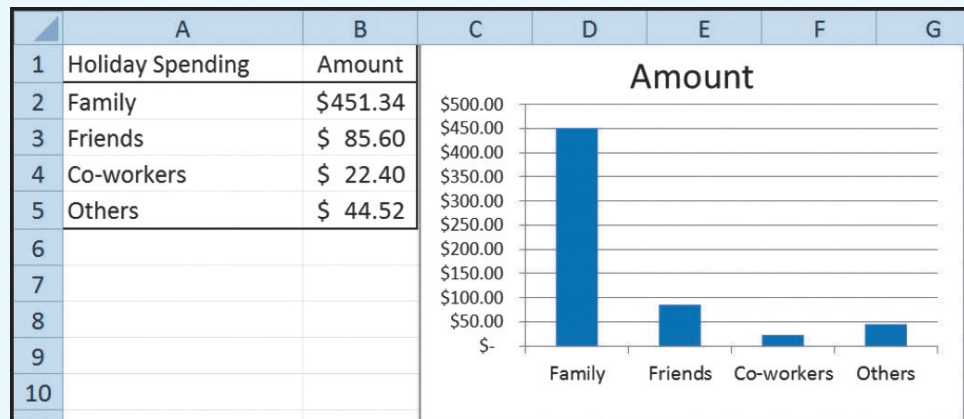
|          | Discrete | Continuous |
|----------|----------|------------|
| Nominal  |          |            |
| Ordinal  |          |            |
| Interval |          |            |
| Ratio    |          | a. Salary  |

14. Using data from such publications as the *Statistical Abstract of the United States*, *Forbes*, or any news source, give examples of variables measured with nominal, ordinal, interval, and ratio scales.
15. The Struthers Wells Corporation employs more than 10,000 white-collar workers in its sales offices and manufacturing facilities in the United States, Europe, and Asia. A sample of 300 U.S. workers revealed 120 would accept a transfer to a location outside the United States. On the basis of these findings, write a brief memo to Ms. Wanda Carter, Vice President of Human Services, regarding all white-collar workers in the firm and their willingness to relocate.
16. AVX Home Entertainment, Inc., recently began a “no-hassles” return policy. A sample of 500 customers who recently returned items showed 400 thought the policy was fair, 32 thought it took too long to complete the transaction, and the rest had no opinion. On the basis of this information, make an inference about customer reaction to the new policy.
17. **FILE** *The Wall Street Journal's* website, [www.wsj.com](http://www.wsj.com), reported the total number of cars and light-duty trucks sold through February of 2016 and February of 2017. The top 16 of 29 manufacturers are listed here. Sales data often are reported in this way to compare current sales to last year’s sales. See the data file for the complete list and use it to respond to the following questions.



| Manufacturer                 | Year-to-Date Sales    |                       |
|------------------------------|-----------------------|-----------------------|
|                              | Through February 2017 | Through February 2016 |
| General Motors Corp.         | 433,049               | 431,570               |
| Ford Motor Company           | 378,650               | 388,523               |
| Toyota Motor Sales USA Inc.  | 317,387               | 349,238               |
| Chrysler                     | 315,684               | 353,525               |
| Nissan North America Inc.    | 248,059               | 236,645               |
| American Honda Motor Co Inc. | 228,066               | 219,482               |
| Hyundai Motor America        | 99,527                | 98,020                |
| Subaru of America Inc.       | 89,379                | 83,112                |
| Kia Motors America Inc.      | 78,299                | 88,042                |
| Mercedes-Benz                | 54,611                | 51,773                |
| Volkswagen of America Inc.   | 48,655                | 42,400                |
| Mazda Motor of America Inc.  | 44,522                | 41,247                |
| BMW of North America Inc.    | 40,667                | 40,580                |
| Audi of America Inc.         | 26,942                | 23,568                |
| Mitsubishi Motors N A Inc.   | 17,381                | 14,134                |
| Volvo                        | 8,123                 | 9,504                 |

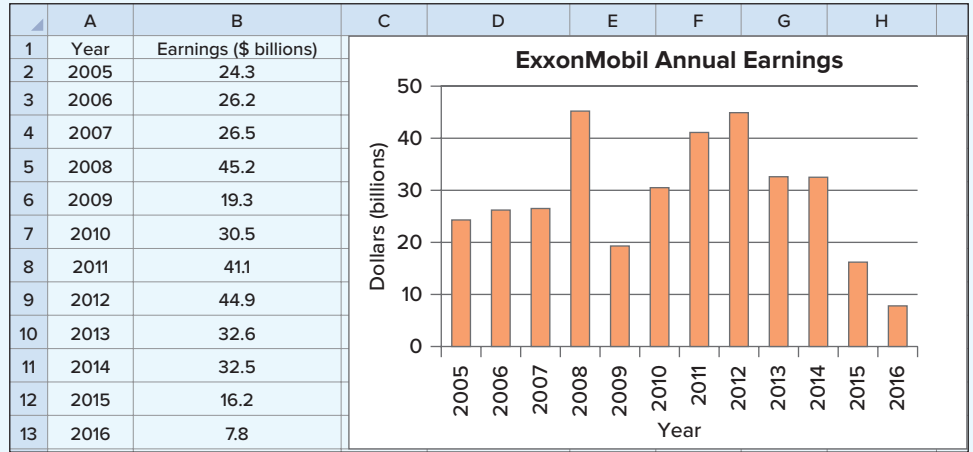
- Using computer software, compare the 2016 year-to-date sales through February to the 2017 year-to-date sales through February for each manufacturer by computing the difference. Using computer software, sort the list of the manufacturers in order of largest to smallest difference in year-to-date sales.
  - For each manufacturer, compare 2016 year-to-date sales through February to 2017 year-to-date sales through February by computing the percentage change in year-to-date sales using computer software. Using computer software, sort the list of manufacturers from largest to smallest using the percentage change. Which manufacturers are in the top five in percentage change? Which manufacturers are in the bottom five in percentage change?
  - Using computer software, sort the list of manufacturers from largest to smallest using 2017 year-to-date sales through February. Then, design a bar graph to illustrate the 2016 and 2017 year-to-date sales through February for the top 12 manufacturers. Also, design a bar graph to illustrate the percentage change for the top 12 manufacturers. Compare these two graphs and prepare brief written comments.
- 18.** The following chart depicts the average amounts spent by consumers on holiday gifts.



Source: Microsoft Excel

Write a brief report summarizing the amounts spent during the holidays. Be sure to include the total amount spent and the percent spent by each group.

19. The following chart depicts the earnings in billions of dollars for ExxonMobil for the period 2005 until 2016. Write a brief report discussing the earnings at ExxonMobil during the period. Was one year higher than the others? Did the earnings increase, decrease, or stay the same over the period?



Source: Microsoft Excel

## DATA ANALYTICS

(The data for these exercises are available in Connect.)

20. **FILE** Refer to the North Valley Real Estate data, which report information on homes sold in the area last year. Consider the following variables: selling price, number of bedrooms, township, and mortgage type.
- Which of the variables are qualitative and which are quantitative?
  - How is each variable measured? Determine the level of measurement for each of the variables.
21. **FILE** Refer to the Baseball 2016 data, which report information on the 30 Major League Baseball teams for the 2016 season. Consider the following variables: number of wins, payroll, season attendance, whether the team is in the American or National League, and the number of home runs hit.
- Which of these variables are quantitative and which are qualitative?
  - Determine the level of measurement for each of the variables.
22. **FILE** Refer to the Lincolnville School District bus data, which report information on the school district’s bus fleet.
- Which of the variables are qualitative and which are quantitative?
  - Determine the level of measurement for each variable.

## PRACTICE TEST

There is a practice test at the end of each chapter. The tests are in two parts. The first part includes 10 to 15 objective questions, usually in a fill-in-the-blank format. The second part includes problems. In most cases, it should take 30 to 45 minutes to complete the test. The problems will require a calculator. Check your answers against those provided in Appendix D in the back of the book.

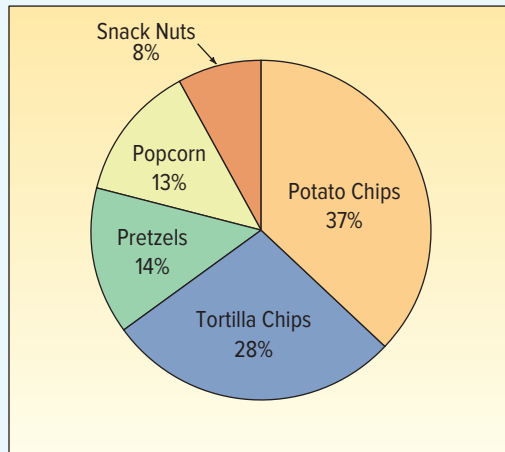
### Part 1—Objective

- The science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions is referred to as \_\_\_\_\_.
- Methods of organizing, summarizing, and presenting data in an enlightening way are called \_\_\_\_\_.
- The methods used to estimate a value of a population on the basis of a sample are called \_\_\_\_\_.
- A portion, or part, of the group of interest is referred to as a \_\_\_\_\_.
- The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest is known as a \_\_\_\_\_.
- With the \_\_\_\_\_ level of measurement, the data are sorted into categories with no particular order to the categories.

7. The \_\_\_\_\_ level of measurement has a significant zero point.
8. The \_\_\_\_\_ level of measurement presumes that one classification is ranked higher than another.
9. The \_\_\_\_\_ level of measurement has the characteristic that the distance between values is a constant size.
10. Is the number of bedrooms in a house a discrete or continuous variable? \_\_\_\_\_
11. The jersey numbers on baseball uniforms are an example of the \_\_\_\_\_ level of measurement.
12. What level of measurement is used when students are classified by eye color? \_\_\_\_\_

### Part 2—Problems

1. Thirty million pounds of snack food were eaten during a recent Super Bowl Sunday. The chart below describes this information.



- a. Estimate, in millions of pounds, the amount of potato chips eaten during the game.
  - b. Calculate approximately the ratio of potato chips consumed to popcorn consumed (twice as much, half as much, three times as much, etc.).
  - c. What percent of the total consists of potato chips and tortilla chips?
2. There are 14 freshmen, 18 sophomores, 10 juniors, and 6 seniors enrolled in an introductory finance class. Answer the following questions.
    - a. What is the level of measurement for these student data?
    - b. What percent of the students are either freshmen or sophomores?

# Describing Data:

## FREQUENCY TABLES, FREQUENCY DISTRIBUTIONS, AND GRAPHIC PRESENTATION

# 2



©goodluz/Shutterstock

- ▲ **MERRILL LYNCH** recently completed a study of online investment portfolios for a sample of clients. For the 70 participants in the study, organize these data into a frequency distribution. (See Exercise 43 and **LO2-3**.)

### LEARNING OBJECTIVES

---

*When you have completed this chapter, you will be able to:*

- LO2-1** Summarize qualitative variables with frequency and relative frequency tables.
- LO2-2** Display a frequency table using a bar or pie chart.
- LO2-3** Summarize quantitative variables with frequency and relative frequency distributions.
- LO2-4** Display a frequency distribution using a histogram or frequency polygon.

## INTRODUCTION

The United States automobile retailing industry is highly competitive. It is dominated by megadealerships that own and operate 50 or more franchises, employ over 10,000 people, and generate several billion dollars in annual sales. Many of the top dealerships



©Darren Brode/Shutterstock

are publicly owned, with shares traded on the New York Stock Exchange or NASDAQ. In 2017, the largest megadealership was AutoNation (ticker symbol AN), followed by Penske Auto Group (PAG), Group 1 Automotive, Inc. (ticker symbol GPI), and the privately owned Van Tuyl Group.

These large corporations use statistics and analytics to summarize and analyze data and information to support their decisions. As an example, we will look at the Applewood Auto Group. It owns four dealerships and sells a wide range of vehicles. These include the popular Korean brands Kia and Hyundai, BMW and Volvo sedans and luxury SUVs, and a full line of Ford and Chevrolet cars and trucks.

Ms. Kathryn Ball is a member of the senior management team at Applewood Auto Group, which has its corporate offices adjacent to Kane Motors. She is responsible for tracking and analyzing vehicle sales and the profitability of those vehicles. Kathryn would like to summarize the profit earned on the vehicles sold using tables, charts, and graphs that she would review and present to the ownership group monthly. She wants to know the profit per vehicle sold, as well as the lowest and highest amount of profit. She is also interested in describing the demographics of the buyers. What are their ages? How many vehicles have they previously purchased from one of the Applewood dealerships? What type of vehicle did they purchase?

The Applewood Auto Group operates four dealerships:

- **Tionesta Ford Lincoln** sells Ford and Lincoln cars and trucks.
- **Olean Automotive Inc.** has the Nissan franchise as well as the General Motors brands of Chevrolet, Cadillac, and GMC trucks.
- **Sheffield Motors Inc.** sells Buick, GMC trucks, Hyundai, and Kia.
- **Kane Motors** offers the Chrysler, Dodge, and Jeep lines as well as BMW and Volvo.

| APPLEWOOD AUTO GROUP |     |         |           |              |          |
|----------------------|-----|---------|-----------|--------------|----------|
|                      | A   | B       | C         | D            | E        |
| 1                    | Age | Profit  | Location  | Vehicle-Type | Previous |
| 2                    | 21  | \$1,387 | Tionesta  | Sedan        | 0        |
| 3                    | 23  | \$1,754 | Sheffield | SUV          | 1        |
| 4                    | 24  | \$1,817 | Sheffield | Hybrid       | 1        |
| 5                    | 25  | \$1,040 | Sheffield | Compact      | 0        |
| 6                    | 26  | \$1,273 | Kane      | Sedan        | 1        |
| 7                    | 27  | \$1,529 | Sheffield | Sedan        | 1        |
| 8                    | 27  | \$3,082 | Kane      | Truck        | 0        |
| 9                    | 28  | \$1,951 | Kane      | SUV          | 1        |
| 10                   | 28  | \$2,692 | Tionesta  | Compact      | 0        |
| 11                   | 29  | \$1,206 | Sheffield | Sedan        | 0        |
| 12                   | 29  | \$1,342 | Kane      | Sedan        | 2        |
| 13                   | 30  | \$443   | Kane      | Sedan        | 3        |
| 14                   | 30  | \$754   | Olean     | Sedan        | 2        |
| 15                   | 30  | \$1,621 | Sheffield | Truck        | 1        |

Source: Microsoft Excel

### LO2-1

Summarize qualitative variables with frequency and relative frequency tables.

Every month, Ms. Ball collects data from each of the four dealerships and enters them into an Excel spreadsheet. Last month the Applewood Auto Group sold 180 vehicles at the four dealerships. A copy of the first few observations appears to the left. The variables collected include:

- **Age**—the age of the buyer at the time of the purchase.
- **Profit**—the amount earned by the dealership on the sale of each vehicle.
- **Location**—the dealership where the vehicle was purchased.
- **Vehicle type**—SUV, sedan, compact, hybrid, or truck.
- **Previous**—the number of vehicles previously purchased at any of the four Applewood dealerships by the consumer.

The entire data set is available in Connect and in Appendix A.4 at the end of the text.

## CONSTRUCTING FREQUENCY TABLES

Recall from Chapter 1 that techniques used to describe a set of data are called descriptive statistics. Descriptive statistics organize data to show the general pattern of the data, to identify where values tend to concentrate, and to expose extreme or unusual data values. The first technique we discuss is a **frequency table**.

**FREQUENCY TABLE** A grouping of qualitative data into mutually exclusive and collectively exhaustive classes showing the number of observations in each class.

In Chapter 1, we distinguished between qualitative and quantitative variables. To review, a qualitative variable is nonnumeric, that is, it can only be classified into distinct categories. Examples of qualitative data include political affiliation (Republican, Democrat, Independent, or other), state of birth (Alabama, . . . , Wyoming), and method of payment for a purchase at Barnes & Noble (cash, digital wallet, debit, or credit). On the other hand, quantitative variables are numerical in nature. Examples of quantitative data relating to college students include the price of their textbooks, their age, and the number of credit hours they are registered for this semester.

In the Applewood Auto Group data set, there are five variables for each vehicle sale: age of the buyer, amount of profit, dealer that made the sale, type of vehicle sold, and number of previous purchases by the buyer. The dealer and the type of vehicle are *qualitative* variables. The amount of profit, the age of the buyer, and the number of previous purchases are *quantitative* variables.



©Dragon Images/Shutterstock

Suppose Ms. Ball wants to summarize last month's sales by location. The first step is to sort the vehicles sold last month according to their location and then tally, or count, the number sold at each of the four locations: Tionesta, Olean, Sheffield, or Kane. The four locations are used to develop a frequency table with four mutually exclusive (distinctive) classes. Mutually exclusive classes means that a particular vehicle can be assigned to only one class. In addition, the frequency table must be collectively exhaustive. That is, every vehicle sold last month is accounted for in the table. If every vehicle is included in the frequency table, the table will be collectively exhaustive and the total number of vehicles will be 180. How do we obtain these counts? Excel provides a tool called a Pivot Table that will quickly and accurately establish the four classes and do the counting. The Excel results follow in Table 2–1. The table shows a total of 180 vehicles; of the 180 vehicles, 52 were sold at Kane Motors.

**TABLE 2–1** Frequency Table for Vehicles Sold Last Month at Applewood Auto Group by Location

| Location  | Number of Cars |
|-----------|----------------|
| Kane      | 52             |
| Olean     | 40             |
| Sheffield | 45             |
| Tionesta  | 43             |
| Total     | 180            |

## Relative Class Frequencies

You can convert class frequencies to relative class frequencies to show the fraction of the total number of observations in each class. A relative frequency captures the relationship between a class frequency and the total number of observations. In the vehicle sales example, we may want to know the percentage of total cars sold at each of the four locations. To convert a frequency table to a relative frequency table, each of the class frequencies is divided by the total number of observations. Again, this is easily accomplished using Excel. The fraction of vehicles sold last month at the Kane location is 0.289, found by 52 divided by 180. The relative frequency for each location is shown in Table 2–2.

**TABLE 2–2** Relative Frequency Table of Vehicles Sold by Location Last Month at Applewood Auto Group

| Location  | Number of Cars | Relative Frequency | Found by |
|-----------|----------------|--------------------|----------|
| Kane      | 52             | .289               | 52/180   |
| Olean     | 40             | .222               | 40/180   |
| Sheffield | 45             | .250               | 45/180   |
| Tionesta  | 43             | .239               | 43/180   |
| Total     | 180            | 1.000              |          |

**LO2-2**

Display a frequency table using a bar or pie chart.

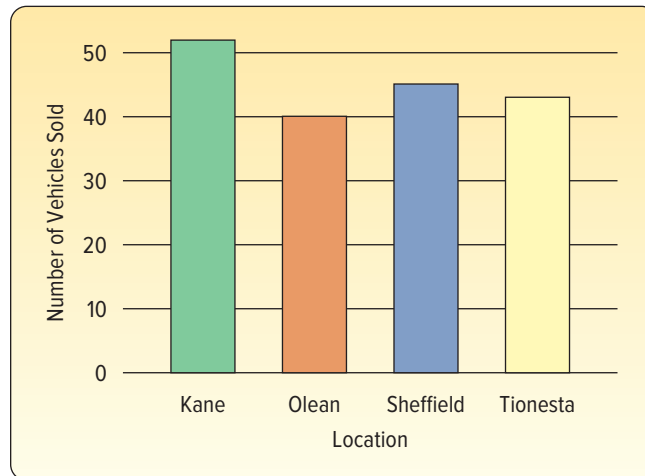
## GRAPHIC PRESENTATION OF QUALITATIVE DATA

The most common graphic form to present a qualitative variable is a **bar chart**. In most cases, the horizontal axis shows the variable of interest. The vertical axis shows the frequency or fraction of each of the possible outcomes. A distinguishing feature of a bar chart is there is distance or a gap between the bars. That is, because the variable of interest is qualitative, the bars are not adjacent to each other. Thus, a bar chart graphically describes a frequency table using a series of uniformly wide rectangles, where the height of each rectangle is the class frequency.

**BAR CHART** A graph that shows qualitative classes on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are proportional to the heights of the bars.

We use the Applewood Auto Group data as an example (Chart 2–1). The variables of interest are the location where the vehicle was sold and the number of vehicles sold at each location. We label the horizontal axis with the four locations and scale the vertical axis with the number sold. The variable location is of nominal scale, so the order of the locations on the horizontal axis does not matter. In Chart 2–1, the locations are listed alphabetically. The locations could also be in order of decreasing or increasing frequencies.

The height of the bars, or rectangles, corresponds to the number of vehicles at each location. There were 52 vehicles sold last month at the Kane location, so the height of the Kane bar is 52; the height of the bar for the Olean location is 40.



**CHART 2–1** Number of Vehicles Sold by Location

Another useful type of chart for depicting qualitative information is a **pie chart**.

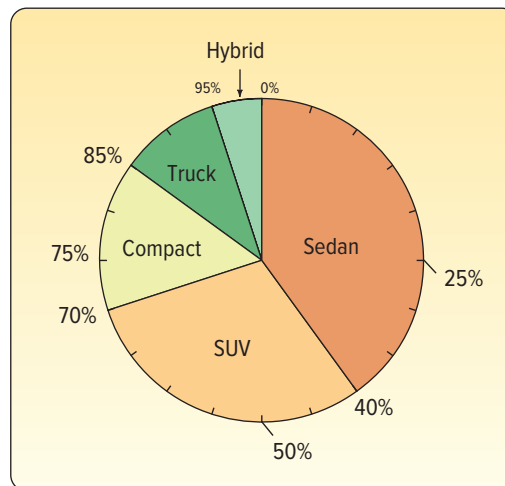
**PIE CHART** A chart that shows the proportion or percentage that each class represents of the total number of frequencies.

We explain the details of constructing a pie chart using the information in Table 2–3, which shows the frequency and percent of cars sold by the Applewood Auto Group for each vehicle type.

**TABLE 2-3** Vehicle Sales by Type at Applewood Auto Group

| Vehicle Type | Number Sold | Percent Sold |
|--------------|-------------|--------------|
| Sedan        | 72          | 40%          |
| SUV          | 54          | 30%          |
| Compact      | 27          | 15%          |
| Truck        | 18          | 10%          |
| Hybrid       | 9           | 5%           |
| Total        | 180         | 100%         |

The first step to develop a pie chart is to mark the percentages 0, 5, 10, 15, and so on evenly around the circumference of a circle (see Chart 2-2). To plot the 40% of total sales represented by sedans, draw a line from the center of the circle to 0 and another line from the center of the circle to 40%. The area in this “slice” represents the number of sedans sold as a percentage of the total sales. Next, add the SUV’s percentage of total sales, 30%, to the sedan’s percentage of total sales, 40%. The result is 70%. Draw a line from the center of the circle to 70%, so the area between 40 and 70 shows the sales of SUVs as a percentage of total sales. Continuing, add the 15% of total sales for compact vehicles, which gives us a total of 85%. Draw a line from the center of the circle to 85, so the “slice” between 70% and 85% represents the number of compact vehicles sold as a percentage of the total sales. The remaining 10% for truck sales and 5% for hybrid sales are added to the chart using the same method.

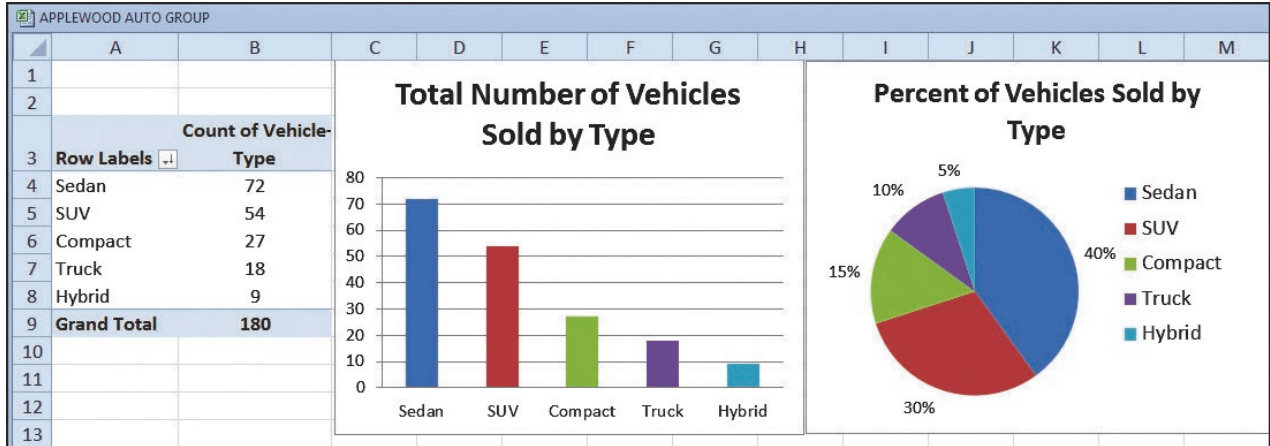
**CHART 2-2** Pie Chart of Vehicles by Type

Because each slice of the pie represents the relative frequency of each vehicle type as a percentage of the total sales, we can easily compare them:

- The largest percentage of sales is for sedans.
- Sedans and SUVs together account for 70% of vehicle sales.
- Hybrids account for 5% of vehicle sales, in spite of being on the market for only a few years.

We can use Excel software to quickly count the number of cars for each vehicle type and create the frequency table, bar chart, and pie chart shown in the following summary. The Excel tool is called a Pivot Table. The instructions to produce these descriptive statistics and charts are given in Appendix C.





Source: Microsoft Excel

Pie and bar charts both serve to illustrate frequency and relative frequency tables. When is a pie chart preferred to a bar chart? In most cases, pie charts are used to show and compare the relative differences in the percentage of observations for each value or class of a qualitative variable. Bar charts are preferred when the goal is to compare the number or frequency of observations for each value or class of a qualitative variable. The following Example/Solution shows another application of bar and pie charts.

**EXAMPLE**

SkiLodges.com is test-marketing its new website and is interested in how easy its website design is to navigate. It randomly selected 200 regular Internet users and asked them to perform a search task on the website. Each person was asked to rate the relative ease of navigation as poor, good, excellent, or awesome. The results are shown in the following table:

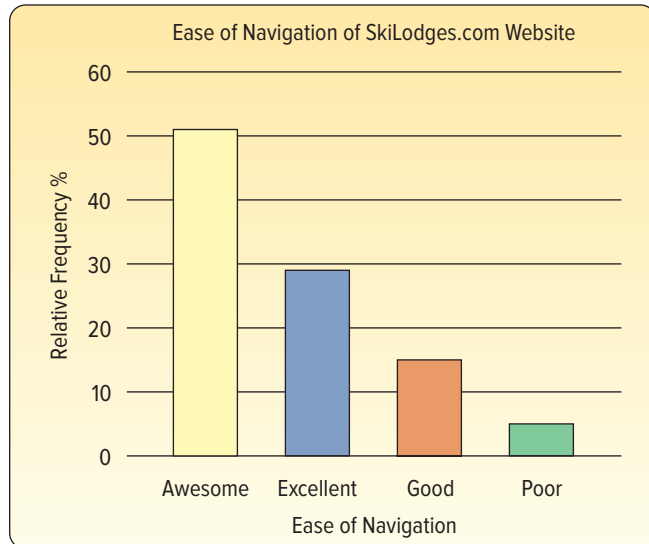
|           |     |
|-----------|-----|
| Awesome   | 102 |
| Excellent | 58  |
| Good      | 30  |
| Poor      | 10  |

1. What type of measurement scale is used for ease of navigation?
2. Draw a bar chart for the survey results.
3. Draw a pie chart for the survey results.

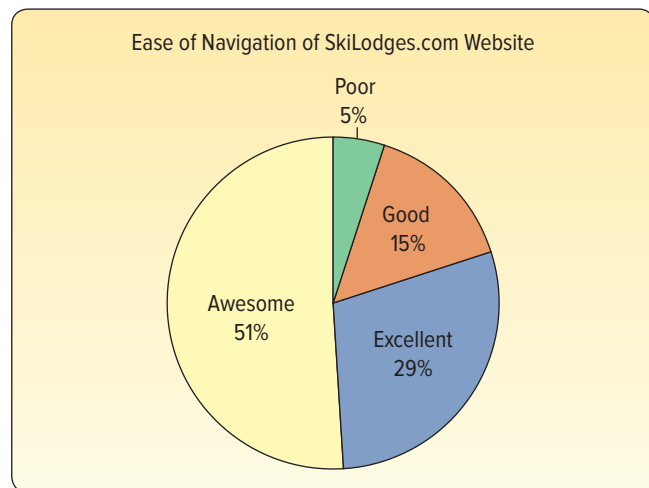
**SOLUTION**

The data are measured on an ordinal scale. That is, the scale is ranked in relative ease of navigation when moving from “awesome” to “poor.” The interval between each rating is unknown so it is impossible, for example, to conclude that a rating of good is twice the value of a poor rating.

We can use a bar chart to graph the data. The vertical scale shows the relative frequency and the horizontal scale shows the values of the ease-of-navigation variable.



A pie chart can also be used to graph these data. The pie chart emphasizes that more than half of the respondents rate the relative ease of using the website as awesome.



## SELF-REVIEW 2-1



The answers are in Appendix E.

DeCenzo Specialty Food and Beverage Company has been serving a cola drink with an additional flavoring, Cola-Plus, that is very popular among its customers. The company is interested in customer preferences for Cola-Plus versus Coca-Cola, Pepsi, and a lemon-lime beverage. They ask 100 randomly sampled customers to take a taste test and select the beverage they prefer most. The results are shown in the following table:

| Beverage   | Number |
|------------|--------|
| Cola-Plus  | 40     |
| Coca-Cola  | 25     |
| Pepsi      | 20     |
| Lemon-Lime | 15     |
| Total      | 100    |

- Are the data qualitative or quantitative? Why?
- What is the table called? What does it show?
- Develop a bar chart to depict the information.
- Develop a pie chart using the relative frequencies.

## EXERCISES

The answers to the odd-numbered exercises are at the end of the book in Appendix D.

- A pie chart shows the relative market share of cola products. The “slice” for Pepsi-Cola has a central angle of 90 degrees. What is its market share?
- In a marketing study, 100 consumers were asked to select the best digital music player from the iPod Touch, Sony Walkman, and Zune HD. To summarize the consumer responses with a frequency table, how many classes would the frequency table have?
- A total of 1,000 residents in Minnesota were asked which season they preferred. One hundred liked winter best, 300 liked spring, 400 liked summer, and 200 liked fall. Develop a frequency table and a relative frequency table to summarize this information.
- Two thousand frequent business travelers were asked which Midwestern city they prefer: Indianapolis, Saint Louis, Chicago, or Milwaukee. One hundred liked Indianapolis best, 450 liked Saint Louis, 1,300 liked Chicago, and the remainder preferred Milwaukee. Develop a frequency table and a relative frequency table to summarize this information.
- Wellstone Inc. produces and markets replacement covers for cell phones in five different colors: bright white, metallic black, magnetic lime, tangerine orange, and fusion red. To estimate the demand for each color, the company set up a kiosk in the Mall of America for several hours and asked randomly selected people which cover color was their favorite. The results follow:

|                  |     |
|------------------|-----|
| Bright white     | 130 |
| Metallic black   | 104 |
| Magnetic lime    | 325 |
| Tangerine orange | 455 |
| Fusion red       | 286 |

- What is the table called?
  - Draw a bar chart for the table.
  - Draw a pie chart.
  - If Wellstone Inc. plans to produce 1 million cell phone covers, how many of each color should it produce?
- A small-business consultant is investigating the performance of several companies. The fourth-quarter sales for last year (in thousands of dollars) for the selected companies were:

| Company                              | Fourth-Quarter Sales<br>(\$ thousands) |
|--------------------------------------|--|
| Hoden Building Products              | \$ 1,645.2                             |
| J & R Printing Inc.                  | 4,757.0                                |
| Long Bay Concrete Construction       | 8,913.0                                |
| Mancell Electric and Plumbing        | 627.1                                  |
| Maxwell Heating and Air Conditioning | 24,612.0                               |
| Mizelle Roofing & Sheet Metals       | 191.9                                  |

The consultant wants to include a chart in his report comparing the sales of the six companies. Use a bar chart to compare the fourth-quarter sales of these corporations and write a brief report summarizing the bar chart.

**LO2-3**

Summarize quantitative variables with frequency and relative frequency distributions.

## CONSTRUCTING FREQUENCY DISTRIBUTIONS

In Chapter 1 and earlier in this chapter, we distinguished between qualitative and quantitative data. In the previous section, using the Applewood Automotive Group data, we summarized two qualitative variables: the location of the sale and the type of vehicle sold. We created frequency and relative frequency tables and depicted the results in bar and pie charts.

The Applewood Auto Group data also include several quantitative variables: the age of the buyer, the profit earned on the sale of the vehicle, and the number of previous purchases. Suppose Ms. Ball wants to summarize last month’s sales by profit earned for each vehicle. We can describe profit using a **frequency distribution**.

**FREQUENCY DISTRIBUTION** A grouping of quantitative data into mutually exclusive and collectively exhaustive classes showing the number of observations in each class.

How do we develop a frequency distribution? The following example shows the steps to construct a frequency distribution. Remember, our goal is to construct tables, charts, and graphs that will quickly summarize the data by showing the location, extreme values, and shape of the data’s distribution.

**EXAMPLE**

Ms. Kathryn Ball of the Applewood Auto Group wants to summarize the quantitative variable profit with a frequency distribution and display the distribution with charts and graphs. With this information, Ms. Ball can easily answer the following questions: What is the typical profit on each sale? What is the largest or maximum profit on any sale? What is the smallest or minimum profit on any sale? Around what value do the profits tend to cluster?

**SOLUTION**

To begin, we show the profits for each of the 180 vehicle sales listed in Table 2–4. This information is called raw or ungrouped data because it is simply a listing

**TABLE 2–4** Profit on Vehicles Sold Last Month by the Applewood Auto Group

|         |         |         |        |        |         |         |         |         |
|---------|---------|---------|--------|--------|---------|---------|---------|---------|
| \$1,387 | \$2,148 | \$2,201 | \$ 963 | \$ 820 | \$2,230 | \$3,043 | \$2,584 | \$2,370 |
| 1,754   | 2,207   | 996     | 1,298  | 1,266  | 2,341   | 1,059   | 2,666   | 2,637   |
| 1,817   | 2,252   | 2,813   | 1,410  | 1,741  | 3,292   | 1,674   | 2,991   | 1,426   |
| 1,040   | 1,428   | 323     | 1,553  | 1,772  | 1,108   | 1,807   | 934     | 2,944   |
| 1,273   | 1,889   | 352     | 1,648  | 1,932  | 1,295   | 2,056   | 2,063   | 2,147   |
| 1,529   | 1,166   | 482     | 2,071  | 2,350  | 1,344   | 2,236   | 2,083   | 1,973   |
| 3,082   | 1,320   | 1,144   | 2,116  | 2,422  | 1,906   | 2,928   | 2,856   | 2,502   |
| 1,951   | 2,265   | 1,485   | 1,500  | 2,446  | 1,952   | 1,269   | 2,989   | 783     |
| 2,692   | 1,323   | 1,509   | 1,549  | 369    | 2,070   | 1,717   | 910     | 1,538   |
| 1,206   | 1,760   | 1,638   | 2,348  | 978    | 2,454   | 1,797   | 1,536   | 2,339   |
| 1,342   | 1,919   | 1,961   | 2,498  | 1,238  | 1,606   | 1,955   | 1,957   | 2,700   |
| 443     | 2,357   | 2,127   | 294    | 1,818  | 1,680   | 2,199   | 2,240   | 2,222   |
| 754     | 2,866   | 2,430   | 1,115  | 1,824  | 1,827   | 2,482   | 2,695   | 2,597   |
| 1,621   | 732     | 1,704   | 1,124  | 1,907  | 1,915   | 2,701   | 1,325   | 2,742   |
| 870     | 1,464   | 1,876   | 1,532  | 1,938  | 2,084   | 3,210   | 2,250   | 1,837   |
| 1,174   | 1,626   | 2,010   | 1,688  | 1,940  | 2,639   | 377     | 2,279   | 2,842   |
| 1,412   | 1,762   | 2,165   | 1,822  | 2,197  | 842     | 1,220   | 2,626   | 2,434   |
| 1,809   | 1,915   | 2,231   | 1,897  | 2,646  | 1,963   | 1,401   | 1,501   | 1,640   |
| 2,415   | 2,119   | 2,389   | 2,445  | 1,461  | 2,059   | 2,175   | 1,752   | 1,821   |
| 1,546   | 1,766   | 335     | 2,886  | 1,731  | 2,338   | 1,118   | 2,058   | 2,487   |

Maximum

Minimum

of the individual, observed profits. It is possible to search the list and find the smallest or minimum profit (\$294) and the largest or maximum profit (\$3,292), but that is about all. It is difficult to determine a typical profit or to visualize where the profits tend to cluster. The raw data are more easily interpreted if we summarize the data with a frequency distribution. The steps to create this frequency distribution follow.

**Step 1: Decide on the number of classes.** A useful recipe to determine the number of classes ( $k$ ) is the “2 to the  $k$  rule.” This guide suggests you select the smallest number ( $k$ ) for the number of classes such that  $2^k$  (in words, 2 raised to the power of  $k$ ) is greater than the number of observations ( $n$ ). In the Applewood Auto Group example, there were 180 vehicles sold. So  $n = 180$ . If we try  $k = 7$ , which means we would use 7 classes,  $2^7 = 128$ , which is less than 180. Hence, 7 is too few classes. If we let  $k = 8$ , then  $2^8 = 256$ , which is greater than 180. So the recommended number of classes is 8.

**Step 2: Determine the class interval.** Generally, the **class interval** is the same for all classes. The classes all taken together must cover at least the distance from the minimum value in the data up to the maximum value. Expressing these words in a formula:

$$i \geq \frac{\text{Maximum Value} - \text{Minimum Value}}{k}$$

where  $i$  is the class interval, and  $k$  is the number of classes.

For the Applewood Auto Group, the minimum value is \$294 and the maximum value is \$3,292. If we need 8 classes, the interval should be:

$$i \geq \frac{\text{Maximum Value} - \text{Minimum Value}}{k} = \frac{\$3,292 - \$294}{8} = \$374.75$$

In practice, this interval size is usually rounded up to some convenient number, such as a multiple of 10 or 100. The value of \$400 is a reasonable choice.

**Step 3: Set the individual class limits.** State clear class limits so you can put each observation into only one category. This means you must avoid overlapping or unclear class limits. For example, classes such as “\$1,300–\$1,400” and “\$1,400–\$1,500” should not be used because it is not clear whether the value \$1,400 is in the first or second class. In this text, we will generally use the format \$1,300 **up to** \$1,400 and \$1,400 **up to** \$1,500 and so on. With this format, it is clear that \$1,399 goes into the first class and \$1,400 in the second.

Because we always round the class interval up to get a convenient class size, we cover a larger than necessary range. For example, using 8 classes with an interval of \$400 in the Applewood Auto Group example results in a range of  $8(\$400) = \$3,200$ . The actual range is \$2,998, found by  $(\$3,292 - \$294)$ . Comparing that value to \$3,200, we have an excess of \$202. Because we need to cover only the range (*Maximum* – *Minimum*), it is natural to put approximately equal amounts of the excess in each of the two tails. Of course, we also should select convenient class limits. A guideline is to make the lower limit of the first class a multiple of the class interval. Sometimes this is not possible, but the lower limit should at least be rounded. So here are the classes we could use for these data.

| Classes             |
|---------------------|
| \$ 200 up to \$ 600 |
| 600 up to 1,000     |
| 1,000 up to 1,400   |
| 1,400 up to 1,800   |
| 1,800 up to 2,200   |
| 2,200 up to 2,600   |
| 2,600 up to 3,000   |
| 3,000 up to 3,400   |

**Step 4: Tally the vehicle profit into the classes and determine the number of observations in each class.** To begin, the profit from the sale of the first vehicle in Table 2–4 is \$1,387. It is tallied in the \$1,000 up to \$1,400 class. The second profit in the first row of Table 2–4 is \$2,148. It is tallied in the \$1,800 up to \$2,200 class. The other profits are tallied in a similar manner. When all the profits are tallied, the table would appear as:

| Profit              | Frequency                            |
|---------------------|--------------------------------------|
| \$ 200 up to \$ 600 | IIII III                             |
| 600 up to 1,000     | IIII III I                           |
| 1,000 up to 1,400   | IIII III III III                     |
| 1,400 up to 1,800   | IIII III III III III III III         |
| 1,800 up to 2,200   | IIII III III III III III III III III |
| 2,200 up to 2,600   | IIII III III III III II              |
| 2,600 up to 3,000   | IIII III III III                     |
| 3,000 up to 3,400   | IIII                                 |

The number of observations in each class is called the **class frequency**. In the \$200 up to \$600 class there are 8 observations, and in the \$600 up to \$1,000 class there are 11 observations. Therefore, the class frequency in the first class is 8 and the class frequency in the second class is 11. There are a total of 180 observations in the entire set of data. So the sum of all the frequencies should be equal to 180. The results of the frequency distribution are in Table 2–5.

**TABLE 2–5** Frequency Distribution of Profit for Vehicles Sold Last Month at Applewood Auto Group

| Profit              | Frequency |
|---------------------|-----------|
| \$ 200 up to \$ 600 | 8         |
| 600 up to 1,000     | 11        |
| 1,000 up to 1,400   | 23        |
| 1,400 up to 1,800   | 38        |
| 1,800 up to 2,200   | 45        |
| 2,200 up to 2,600   | 32        |
| 2,600 up to 3,000   | 19        |
| 3,000 up to 3,400   | 4         |
| Total               | 180       |

Now that we have organized the data into a frequency distribution (see Table 2–5), we can summarize the profits of the vehicles for the Applewood Auto Group. Observe the following:

1. The profits from vehicle sales range between \$200 and \$3,400.
2. The vehicle profits are classified using a class interval of \$400. The class interval is determined by subtracting consecutive lower or upper class limits. For

example, the lower limit of the first class is \$200, and the lower limit of the second class is \$600. The difference is the class interval of \$400.

3. The profits are concentrated between \$1,000 and \$3,000. The profit on 157 vehicles, or 87%, was within this range.
4. For each class, we can determine the typical profit or **class midpoint**. It is half-way between the lower or upper limits of two consecutive classes. It is computed by adding the lower or upper limits of consecutive classes and dividing by 2. Referring to Table 2–5, the lower class limit of the first class is \$200, and the next class limit is \$600. The class midpoint is \$400, found by  $(\$600 + \$200)/2$ . The midpoint best represents, or is typical of, the profits of the vehicles in that class. Applewood sold 8 vehicles with a typical profit of \$400.
5. The largest concentration, or highest frequency, of vehicles sold is in the \$1,800 up to \$2,200 class. There are 45 vehicles in this class. The class midpoint is \$2,000. So we say that the typical profit in the class with the highest frequency is \$2,000.

By using a frequency distribution, Ms. Ball can make a clear presentation and summary of last month's profits.

We admit that arranging the information on profits into a frequency distribution does result in the loss of some detailed information. That is, by organizing the data into a frequency distribution, we cannot pinpoint the exact profit on any vehicle, such as \$1,387, \$2,148, or \$2,201. Further, we cannot tell that the actual minimum profit for any vehicle sold is \$294 or that the maximum profit was \$3,292. However, the lower limit of the first class and the upper limit of the last class convey essentially the same meaning. Likely, Ms. Ball will make the same judgment if she knows the smallest profit is about \$200 that she will if she knows the exact profit is \$292. The advantages of summarizing the 180 profits into a more understandable and organized form more than offset this disadvantage.

When we summarize raw data with frequency distributions, equal class intervals are preferred. However, in certain situations unequal class intervals may be necessary to avoid a large number of classes with very small frequencies. Such is the case in Table 2–6. The U.S. Internal Revenue Service uses unequal-sized class intervals for adjusted gross income on individual tax returns to summarize the number of individual tax returns. If we use our method to find equal class intervals, the  $2^k$  rule results in 25 classes, and

#### STATISTICS IN ACTION

In 1788, James Madison, John Jay, and Alexander Hamilton anonymously published a series of essays entitled *The Federalist*. These Federalist papers were an attempt to convince the people of New York that they should ratify the Constitution. In the course of history, the authorship of most of these papers became known, but 12 remained contested. Through the use of statistical analysis, and particularly studying the frequency distributions of various words, we can now conclude that James Madison is the likely author of the 12 papers. In fact, the statistical evidence that Madison is the author is overwhelming.

**TABLE 2–6** Adjusted Gross Income for Individuals Filing Income Tax Returns

| Adjusted Gross Income    |                            | Number of Returns<br>(in thousands) |
|--------------------------|----------------------------|-------------------------------------|
| No adjusted gross income |                            | 178.2                               |
| \$                       | 1 up to \$ 5,000           | 1,204.6                             |
|                          | 5,000 up to 10,000         | 2,595.5                             |
|                          | 10,000 up to 15,000        | 3,142.0                             |
|                          | 15,000 up to 20,000        | 3,191.7                             |
|                          | 20,000 up to 25,000        | 2,501.4                             |
|                          | 25,000 up to 30,000        | 1,901.6                             |
|                          | 30,000 up to 40,000        | 2,502.3                             |
|                          | 40,000 up to 50,000        | 1,426.8                             |
|                          | 50,000 up to 75,000        | 1,476.3                             |
|                          | 75,000 up to 100,000       | 338.8                               |
|                          | 100,000 up to 200,000      | 223.3                               |
|                          | 200,000 up to 500,000      | 55.2                                |
|                          | 500,000 up to 1,000,000    | 12.0                                |
|                          | 1,000,000 up to 2,000,000  | 5.1                                 |
|                          | 2,000,000 up to 10,000,000 | 3.4                                 |
|                          | 10,000,000 or more         | 0.6                                 |

a class interval of \$400,000, assuming \$0 and \$10,000,000 as the minimum and maximum values for adjusted gross income. Using equal class intervals, the first 13 classes in Table 2–6 would be combined into one class of about 99.9% of all tax returns and 24 classes for the 0.1% of the returns with an adjusted gross income above \$400,000. Using equal class intervals does not provide a good understanding of the raw data. In this case, good judgment in the use of unequal class intervals, as demonstrated in Table 2–6, is required to show the distribution of the number of tax returns filed, especially for incomes under \$500,000.

## SELF-REVIEW 2-2



In the first quarter of last year, the 11 members of the sales staff at Master Chemical Company earned the following commissions:

\$1,650 \$1,475 \$1,510 \$1,670 \$1,595 \$1,760 \$1,540 \$1,495 \$1,590 \$1,625 \$1,510

- What are the values such as \$1,650 and \$1,475 called?
- Using \$1,400 up to \$1,500 as the first class, \$1,500 up to \$1,600 as the second class, and so forth, organize the quarterly commissions into a frequency distribution.
- What are the numbers in the right column of your frequency distribution called?
- Using the frequency distribution of quarterly commissions: What is the class with the highest frequency of earned commissions? What is the smallest commission? What is the largest commission? What is the typical earned commission?

## Relative Frequency Distribution

It may be desirable, as we did earlier with qualitative data, to convert class frequencies to relative class frequencies to show the proportion of the total number of observations in each class. In our vehicle profits, we may want to know what percentage of the vehicle profits are in the \$1,000 up to \$1,400 class. To convert a frequency distribution to a *relative* frequency distribution, each of the class frequencies is divided by the total number of observations. From the distribution of vehicle profits, Table 2–5, the relative frequency for the \$1,000 up to \$1,400 class is 0.128, found by dividing 23 by 180. That is, profit on 12.8% of the vehicles sold is between \$1,000 and \$1,400. The relative frequencies for the remaining classes are shown in Table 2–7.

**TABLE 2-7** Relative Frequency Distribution of Profit for Vehicles Sold Last Month at Applewood Auto Group

| Profit              | Frequency | Relative Frequency | Found by |
|---------------------|-----------|--------------------|----------|
| \$ 200 up to \$ 600 | 8         | .044               | 8/180    |
| 600 up to 1,000     | 11        | .061               | 11/180   |
| 1,000 up to 1,400   | 23        | .128               | 23/180   |
| 1,400 up to 1,800   | 38        | .211               | 38/180   |
| 1,800 up to 2,200   | 45        | .250               | 45/180   |
| 2,200 up to 2,600   | 32        | .178               | 32/180   |
| 2,600 up to 3,000   | 19        | .106               | 19/180   |
| 3,000 up to 3,400   | 4         | .022               | 4/180    |
| Total               | 180       | 1.000              |          |

| APPLEWOOD AUTO GROUP |              |           |                    |
|----------------------|--------------|-----------|--------------------|
|                      | A            | B         | C                  |
| 1                    | Profit Class | Frequency | Relative Frequency |
| 2                    | 200-600      | 8         | 4.44%              |
| 3                    | 600-1000     | 11        | 6.11%              |
| 4                    | 1000-1400    | 23        | 12.78%             |
| 5                    | 1400-1800    | 38        | 21.11%             |
| 6                    | 1800-2200    | 45        | 25.00%             |
| 7                    | 2200-2600    | 32        | 17.78%             |
| 8                    | 2600-3000    | 19        | 10.56%             |
| 9                    | 3000-3400    | 4         | 2.22%              |
| 10                   | Grand Total  | 180       | 100.00%            |

There are many software packages that perform statistical calculations. Throughout this text, we will show the output from Microsoft Excel, MegaStat (a Microsoft Excel add-in), and Minitab (a statistical software package). Because Excel is most readily available, it is used most frequently.

Within the earlier Graphic Presentation of Qualitative Data section, we used the Pivot Table tool in Excel to create a frequency table. To create the table to the left, we use the same Excel tool to

Source: Microsoft Excel



compute frequency and relative frequency distributions for the profit variable in the Applewood Auto Group data. The necessary steps are given in the Software Commands section in Appendix C.

## SELF-REVIEW 2-3



Barry Bonds of the San Francisco Giants established a new single-season Major League Baseball home run record by hitting 73 home runs during the 2001 season. This record still stands today. Listed below is the sorted distance of each of the 73 home runs.

|     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 320 | 320 | 347 | 350 | 360 | 360 | 360 | 361 | 365 | 370 |
| 370 | 375 | 375 | 375 | 375 | 380 | 380 | 380 | 380 | 380 |
| 380 | 390 | 390 | 391 | 394 | 396 | 400 | 400 | 400 | 400 |
| 405 | 410 | 410 | 410 | 410 | 410 | 410 | 410 | 410 | 410 |
| 410 | 410 | 411 | 415 | 415 | 416 | 417 | 417 | 420 | 420 |
| 420 | 420 | 420 | 420 | 420 | 420 | 429 | 430 | 430 | 430 |
| 430 | 430 | 435 | 435 | 436 | 440 | 440 | 440 | 440 | 440 |
| 450 | 480 | 488 |     |     |     |     |     |     |     |

- For these data, show that seven classes would be used to create a frequency distribution using the  $2^k$  rule.
- Show that a class interval of 30 would summarize the data in seven classes.
- Construct frequency and relative frequency distributions for the data with seven classes and a class interval of 30. Start the first class with a lower limit of 300.
- How many home runs traveled a distance of 360 up to 390 feet?
- What percentage of the home runs traveled a distance of 360 up to 390 feet?
- What percentage of the home runs traveled a distance of 390 feet or more?

## EXERCISES

This **FILE** icon indicates that the data are available in Connect. You will be able to download the data directly into Excel or Minitab from this site.

- A set of data consists of 38 observations. How many classes would you recommend for the frequency distribution?
- A set of data consists of 45 observations between \$0 and \$29. What size would you recommend for the class interval?
- A set of data consists of 230 observations between \$235 and \$567. What class interval would you recommend?
- A set of data contains 53 observations. The minimum value is 42 and the maximum value is 129. The data are to be organized into a frequency distribution.
  - How many classes would you suggest?
  - What would you suggest as the lower limit of the first class?
- FILE** Wachesaw Manufacturing Inc. produced the following number of units in the last 16 days.

|    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|
| 27 | 27 | 27 | 28 | 27 | 25 | 25 | 28 |
| 26 | 28 | 26 | 28 | 31 | 30 | 26 | 26 |

The information is to be organized into a frequency distribution.

- How many classes would you recommend?
- What class interval would you suggest?
- What lower limit would you recommend for the first class?
- Organize the information into a frequency distribution and determine the relative frequency distribution.
- Comment on the shape of the distribution.

12. **FILE** The Quick Change Oil Company has a number of outlets in the metropolitan Seattle area. The daily number of oil changes at the Oak Street outlet in the past 20 days are:

|    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|
| 65 | 98 | 55 | 62 | 79 | 59 | 51 | 90 | 72 | 56 |
| 70 | 62 | 66 | 80 | 94 | 79 | 63 | 73 | 71 | 85 |

The data are to be organized into a frequency distribution.

- How many classes would you recommend?
  - What class interval would you suggest?
  - What lower limit would you recommend for the first class?
  - Organize the number of oil changes into a frequency distribution.
  - Comment on the shape of the frequency distribution. Also, determine the relative frequency distribution.
13. **FILE** The manager of the BiLo Supermarket in Mt. Pleasant, Rhode Island, gathered the following information on the number of times a customer visits the store during a month. The responses of 51 customers were:

|   |    |   |   |   |    |    |   |    |   |   |   |   |    |   |
|---|----|---|---|---|----|----|---|----|---|---|---|---|----|---|
| 5 | 3  | 3 | 1 | 4 | 4  | 5  | 6 | 4  | 2 | 6 | 6 | 6 | 7  | 1 |
| 1 | 14 | 1 | 2 | 4 | 4  | 4  | 5 | 6  | 3 | 5 | 3 | 4 | 5  | 6 |
| 8 | 4  | 7 | 6 | 5 | 9  | 11 | 3 | 12 | 4 | 7 | 6 | 5 | 15 | 1 |
| 1 | 10 | 8 | 9 | 2 | 12 |    |   |    |   |   |   |   |    |   |

- Starting with 0 as the lower limit of the first class and using a class interval of 3, organize the data into a frequency distribution.
  - Describe the distribution. Where do the data tend to cluster?
  - Convert the distribution to a relative frequency distribution.
14. **FILE** The food services division of Cedar River Amusement Park Inc. is studying the amount of money spent per day on food and drink by families who visit the amusement park. A sample of 40 families who visited the park yesterday revealed they spent the following amounts:

|      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| \$77 | \$18 | \$63 | \$84 | \$38 | \$54 | \$50 | \$59 | \$54 | \$56 | \$36 | \$26 | \$50 | \$34 | \$44 |
| 41   | 58   | 58   | 53   | 51   | 62   | 43   | 52   | 53   | 63   | 62   | 62   | 65   | 61   | 52   |
| 60   | 60   | 45   | 66   | 83   | 71   | 63   | 58   | 61   | 71   |      |      |      |      |      |

- Organize the data into a frequency distribution, using seven classes and 15 as the lower limit of the first class. What class interval did you select?
- Where do the data tend to cluster?
- Describe the distribution.
- Determine the relative frequency distribution.

#### LO2-4

Display a frequency distribution using a histogram or frequency polygon.

## GRAPHIC PRESENTATION OF A DISTRIBUTION

Sales managers, stock analysts, hospital administrators, and other busy executives often need a quick picture of the distributions of sales, stock prices, or hospital costs. These distributions can often be depicted by the use of charts and graphs. Three charts that will help portray a frequency distribution graphically are the histogram, the frequency polygon, and the cumulative frequency polygon.

### Histogram

A **histogram** for a frequency distribution based on quantitative data is similar to the bar chart showing the distribution of qualitative data. The classes are marked on the

horizontal axis and the class frequencies on the vertical axis. The class frequencies are represented by the heights of the bars. However, there is one important difference based on the nature of the data. Quantitative data are usually measured using scales that are continuous, not discrete. Therefore, the horizontal axis represents all possible values, and the bars are drawn adjacent to each other to show the continuous nature of the data.

**HISTOGRAM** A graph in which the classes are marked on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are represented by the heights of the bars, and the bars are drawn adjacent to each other.

### EXAMPLE

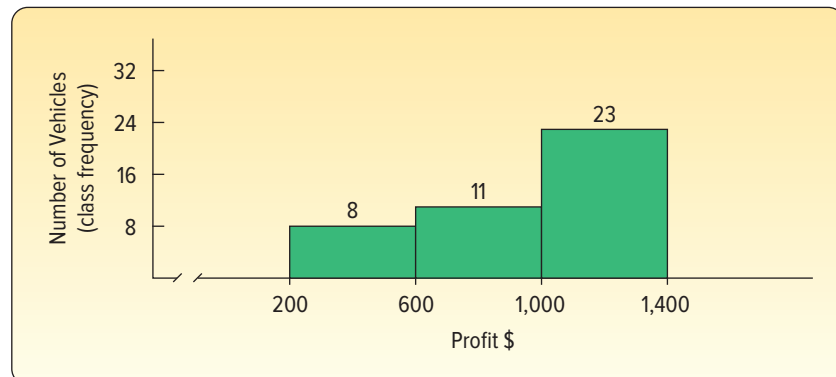
Below is the frequency distribution of the profits on vehicle sales last month at the Applewood Auto Group.

| Profit              | Frequency |
|---------------------|-----------|
| \$ 200 up to \$ 600 | 8         |
| 600 up to 1,000     | 11        |
| 1,000 up to 1,400   | 23        |
| 1,400 up to 1,800   | 38        |
| 1,800 up to 2,200   | 45        |
| 2,200 up to 2,600   | 32        |
| 2,600 up to 3,000   | 19        |
| 3,000 up to 3,400   | 4         |
| Total               | 180       |

Construct a histogram. What observations can you reach based on the information presented in the histogram?

### SOLUTION

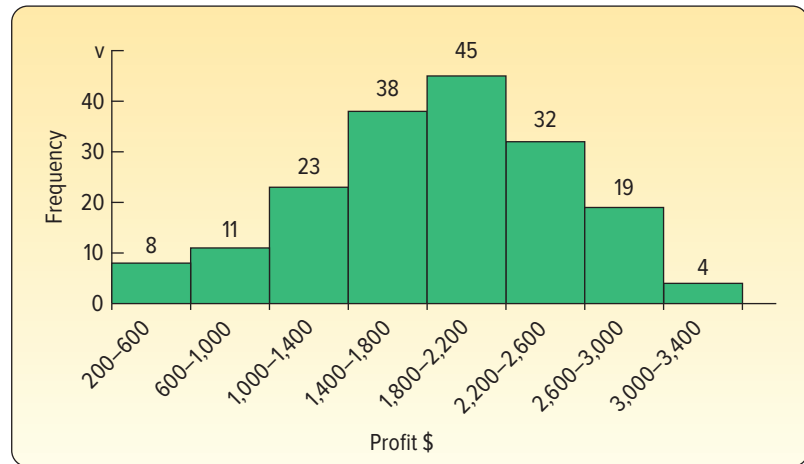
The class frequencies are scaled along the vertical axis (Y-axis) and either the class limits or the class midpoints along the horizontal axis. To illustrate the construction of the histogram, the first three classes are shown in Chart 2–3.



**CHART 2–3** Construction of a Histogram

From Chart 2–3 we note the profit on eight vehicles was \$200 up to \$600. Therefore, the height of the column for that class is 8. There are 11 vehicle sales where the profit was \$600 up to \$1,000. So, logically, the height of that column is 11. The height of the bar represents the number of observations in the class.

This procedure is continued for all classes. The complete histogram is shown in Chart 2–4. Note that there is no space between the bars. This is a feature of the histogram. Why is this so? Because the variable profit, plotted on the horizontal axis, is a continuous variable. In a bar chart, the scale of measurement is usually nominal and the vertical bars are separated. This is an important distinction between the histogram and the bar chart.



**CHART 2–4** Histogram of the Profit on 180 Vehicles Sold at the Applewood Auto Group

We can make the following statements using Chart 2–4. They are the same as the observations based on Table 2–5.

1. The profits from vehicle sales range between \$200 and \$3,400.
2. The vehicle profits are classified using a class interval of \$400. The class interval is determined by subtracting consecutive lower or upper class limits. For example, the lower limit of the first class is \$200, and the lower limit of the second class is \$600. The difference is the class interval or \$400.
3. The profits are concentrated between \$1,000 and \$3,000. The profit on 157 vehicles, or 87%, was within this range.
4. For each class, we can determine the typical profit or class midpoint. It is halfway between the lower or upper limits of two consecutive classes. It is computed by adding the lower or upper limits of consecutive classes and dividing by 2. Referring to Chart 2–4, the lower class limit of the first class is \$200, and the next class limit is \$600. The class midpoint is \$400, found by  $(\$600 + \$200)/2$ . The midpoint best represents, or is typical of, the profits of the vehicles in that class. Applewood sold 8 vehicles with a typical profit of \$400.
5. The largest concentration, or highest frequency of vehicles sold, is in the \$1,800 up to \$2,200 class. There are 45 vehicles in this class. The class midpoint is \$2,000. So we say that the typical profit in the class with the highest frequency is \$2,000.

Thus, the histogram provides an easily interpreted visual representation of a frequency distribution. We should also point out that we would have made the same observations and the shape of the histogram would have been the same had we used a relative frequency distribution instead of the actual frequencies. That is, if we use the relative frequencies of Table 2–7, the result is a histogram of the same shape as Chart 2–4. The only difference is that the vertical axis would have been reported in percentage of vehicles instead of the number of vehicles. The Excel commands to create Chart 2–4 are given in Appendix C.

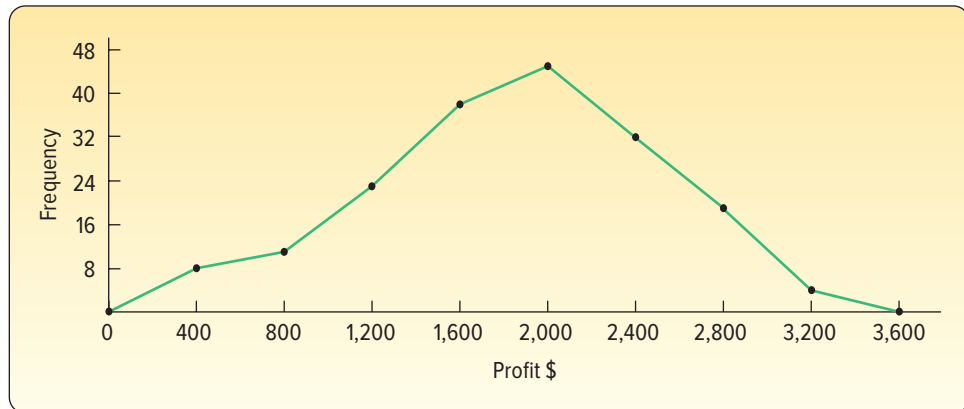
## STATISTICS IN ACTION

Florence Nightingale is known as the founder of the nursing profession. However, she also saved many lives by using statistical analysis. When she encountered an unsanitary condition or an undersupplied hospital, she improved the conditions and then used statistical data to document the improvement. Thus, she was able to convince others of the need for medical reform, particularly in the area of sanitation. She developed original graphs to demonstrate that, during the Crimean War, more soldiers died from unsanitary conditions than were killed in combat.

## Frequency Polygon

A **frequency polygon** also shows the shape of a distribution and is similar to a histogram. It consists of line segments connecting the points formed by the intersections of the class midpoints and the class frequencies. The construction of a frequency polygon is illustrated in Chart 2–5. We use the profits from the cars sold last month at the Applewood Auto Group. The midpoint of each class is scaled on the X-axis and the class frequencies on the Y-axis. Recall that the class midpoint is the value at the center of a class and represents the typical values in that class. The class frequency is the number of observations in a particular class. The profit earned on the vehicles sold last month by the Applewood Auto Group is repeated below.

| Profit              | Midpoint | Frequency |
|---------------------|----------|-----------|
| \$ 200 up to \$ 600 | \$ 400   | 8         |
| 600 up to 1,000     | 800      | 11        |
| 1,000 up to 1,400   | 1,200    | 23        |
| 1,400 up to 1,800   | 1,600    | 38        |
| 1,800 up to 2,200   | 2,000    | 45        |
| 2,200 up to 2,600   | 2,400    | 32        |
| 2,600 up to 3,000   | 2,800    | 19        |
| 3,000 up to 3,400   | 3,200    | 4         |
| Total               |          | 180       |



**CHART 2–5** Frequency Polygon of Profit on 180 Vehicles Sold at Applewood Auto Group

As noted previously, the \$200 up to \$600 class is represented by the midpoint \$400. To construct a frequency polygon, move horizontally on the graph to the midpoint, \$400, and then vertically to 8, the class frequency, and place a dot. The  $x$  and the  $y$  values of this point are called the *coordinates*. The coordinates of the next point are  $x = 800$  and  $y = 11$ . The process is continued for all classes. Then the points are connected in order. That is, the point representing the lowest class is joined to the one representing the second class and so on. Note in Chart 2–5 that, to complete the frequency polygon, midpoints of \$0 and \$3,600 are added to the  $X$ -axis to “anchor” the polygon at zero frequencies. These two values, \$0 and \$3,600, were derived by subtracting the class interval of \$400 from the lowest midpoint (\$400) and by adding \$400 to the highest midpoint (\$3,200) in the frequency distribution.

Both the histogram and the frequency polygon allow us to get a quick picture of the main characteristics of the data (highs, lows, points of concentration, etc.). Although the two representations are similar in purpose, the histogram has the advantage of depicting each class as a rectangle, with the height of the rectangular bar representing



**CHART 2-6** Distribution of Profit at Applewood Auto Group and Fowler Motors

the number in each class. The frequency polygon, in turn, has an advantage over the histogram. It allows us to compare directly two or more frequency distributions. Suppose Ms. Ball wants to compare the profit per vehicle sold at Applewood Auto Group with a similar auto group, Fowler Auto in Grayling, Michigan. To do this, two frequency polygons are constructed, one on top of the other, as in Chart 2–6. Two things are clear from the chart:

- The typical vehicle profit is larger at Fowler Motors—about \$2,000 for Applewood and about \$2,400 for Fowler.
- There is less variation or dispersion in the profits at Fowler Motors than at Applewood. The lower limit of the first class for Applewood is \$0 and the upper limit is \$3,600. For Fowler Motors, the lower limit is \$800 and the upper limit is the same: \$3,600.

The total number of cars sold at the two dealerships is about the same, so a direct comparison is possible. If the difference in the total number of cars sold is large, then converting the frequencies to relative frequencies and then plotting the two distributions would allow a clearer comparison.

### SELF-REVIEW 2-4



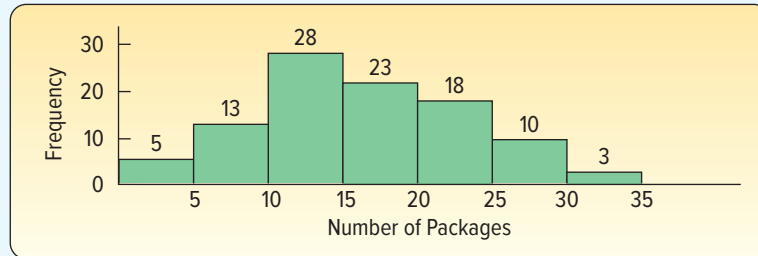
The annual imports of a selected group of electronic suppliers are shown in the following frequency distribution.

| Imports (\$ millions) | Number of Suppliers |
|-----------------------|---------------------|
| \$ 2 up to \$ 5       | 6                   |
| 5 up to 8             | 13                  |
| 8 up to 11            | 20                  |
| 11 up to 14           | 10                  |
| 14 up to 17           | 1                   |

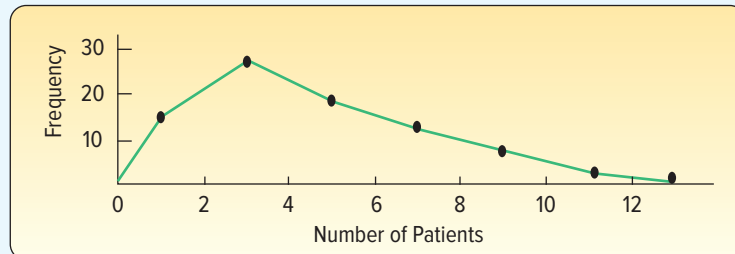
- Portray the imports as a histogram.
- Portray the imports as a relative frequency polygon.
- Summarize the important facets of the distribution (such as classes with the highest and lowest frequencies).

## EXERCISES

15. Molly's Candle Shop has several retail stores in the coastal areas of North and South Carolina. Many of Molly's customers ask her to ship their purchases. The following chart shows the number of packages shipped per day for the last 100 days. For example, the first class shows that there were 5 days when the number of packages shipped was 0 up to 5.



- What is this chart called?
  - What is the total number of packages shipped?
  - What is the class interval?
  - What is the number of packages shipped in the 10 up to 15 class?
  - What is the relative frequency of packages shipped in the 10 up to 15 class?
  - What is the midpoint of the 10 up to 15 class?
  - On how many days were there 25 or more packages shipped?
16. The following chart shows the number of patients admitted daily to Memorial Hospital through the emergency room.



- What is the midpoint of the 2 up to 4 class?
  - On how many days were 2 up to 4 patients admitted?
  - What is the class interval?
  - What is this chart called?
17. The following frequency distribution reports the number of frequent flier miles, reported in thousands, for employees of Brumley Statistical Consulting Inc. during the most recent quarter.

| Frequent Flier Miles (000) | Number of Employees |
|----------------------------|---------------------|
| 0 up to 3                  | 5                   |
| 3 up to 6                  | 12                  |
| 6 up to 9                  | 23                  |
| 9 up to 12                 | 8                   |
| 12 up to 15                | 2                   |
| Total                      | 50                  |

- a. How many employees were studied?
  - b. What is the midpoint of the first class?
  - c. Construct a histogram.
  - d. A frequency polygon is to be drawn. What are the coordinates of the plot for the first class?
  - e. Construct a frequency polygon.
  - f. Interpret the frequent flier miles accumulated using the two charts.
18. A large Internet retailer is studying the lead time (elapsed time between when an order is placed and when it is filled) for a sample of recent orders. The lead times are reported in days.

| Lead Time (days) | Frequency |
|------------------|-----------|
| 0 up to 5        | 6         |
| 5 up to 10       | 7         |
| 10 up to 15      | 12        |
| 15 up to 20      | 8         |
| 20 up to 25      | 7         |
| Total            | 40        |

- a. How many orders were studied?
- b. What is the midpoint of the first class?
- c. What are the coordinates of the first class for a frequency polygon?
- d. Draw a histogram.
- e. Draw a frequency polygon.
- f. Interpret the lead times using the two charts.

## Cumulative Distributions

Consider once again the distribution of the profits on vehicles sold by the Applewood Auto Group. Suppose we are interested in the number of vehicles that sold for a profit of less than \$1,400. These values can be approximated by developing a **cumulative frequency distribution** and portraying it graphically in a **cumulative frequency polygon**. Or, suppose we are interested in the profit earned on the lowest-selling 40% of the vehicles. These values can be approximated by developing a **cumulative relative frequency distribution** and portraying it graphically in a **cumulative relative frequency polygon**.

### ▶ EXAMPLE

The frequency distribution of the profits earned at Applewood Auto Group is repeated from Table 2–5.

| Profit              | Frequency |
|---------------------|-----------|
| \$ 200 up to \$ 600 | 8         |
| 600 up to 1,000     | 11        |
| 1,000 up to 1,400   | 23        |
| 1,400 up to 1,800   | 38        |
| 1,800 up to 2,200   | 45        |
| 2,200 up to 2,600   | 32        |
| 2,600 up to 3,000   | 19        |
| 3,000 up to 3,400   | 4         |
| Total               | 180       |



Construct a cumulative frequency polygon to answer the following question: sixty of the vehicles earned a profit of less than what amount? Construct a cumulative relative frequency polygon to answer this question: seventy-five percent of the vehicles sold earned a profit of less than what amount?

### SOLUTION

As the names imply, a cumulative frequency distribution and a cumulative frequency polygon require *cumulative frequencies*. To construct a cumulative frequency distribution, refer to the preceding table and note that there were eight vehicles in which the profit earned was less than \$600. Those 8 vehicles, plus the 11 in the next higher class, for a total of 19, earned a profit of less than \$1,000. The cumulative frequency for the next higher class is 42, found by  $8 + 11 + 23$ . This process is continued for all the classes. All the vehicles earned a profit of less than \$3,400. (See Table 2–8.)

**TABLE 2–8** Cumulative Frequency Distribution for Profit on Vehicles Sold Last Month at Applewood Auto Group

| Profit           | Cumulative Frequency | Found by                              |
|------------------|----------------------|---------------------------------------|
| Less than \$ 600 | 8                    | 8                                     |
| Less than 1,000  | 19                   | $8 + 11$                              |
| Less than 1,400  | 42                   | $8 + 11 + 23$                         |
| Less than 1,800  | 80                   | $8 + 11 + 23 + 38$                    |
| Less than 2,200  | 125                  | $8 + 11 + 23 + 38 + 45$               |
| Less than 2,600  | 157                  | $8 + 11 + 23 + 38 + 45 + 32$          |
| Less than 3,000  | 176                  | $8 + 11 + 23 + 38 + 45 + 32 + 19$     |
| Less than 3,400  | 180                  | $8 + 11 + 23 + 38 + 45 + 32 + 19 + 4$ |

To construct a cumulative relative frequency distribution, we divide the cumulative frequencies by the total number of observations, 180. As shown in Table 2-9, the cumulative relative frequency of the fourth class is  $80/180 = 44\%$ . This means that 44% of the vehicles sold for less than \$1,800.

**TABLE 2–9** Cumulative Relative Frequency Distribution for Profit on Vehicles Sold Last Month at Applewood Auto Group

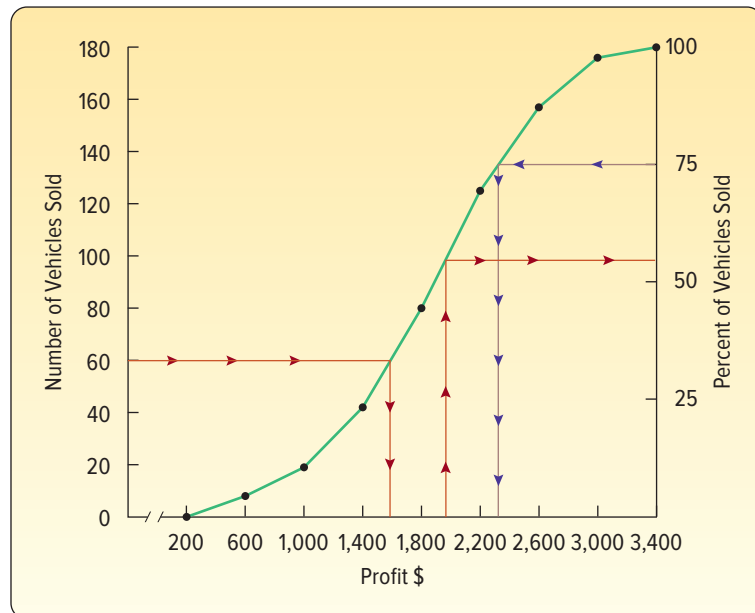
| Profit           | Cumulative Frequency | Cumulative Relative Frequency |
|------------------|----------------------|-------------------------------|
| Less than \$ 600 | 8                    | $8/180 = 0.044 = 4.4\%$       |
| Less than 1,000  | 19                   | $19/180 = 0.106 = 10.6\%$     |
| Less than 1,400  | 42                   | $42/180 = 0.233 = 23.3\%$     |
| Less than 1,800  | 80                   | $80/180 = 0.444 = 44.4\%$     |
| Less than 2,200  | 125                  | $125/180 = 0.694 = 69.4\%$    |
| Less than 2,600  | 157                  | $157/180 = 0.872 = 87.2\%$    |
| Less than 3,000  | 176                  | $176/180 = 0.978 = 97.8\%$    |
| Less than 3,400  | 180                  | $180/180 = 1.000 = 100\%$     |

To plot a cumulative frequency distribution, scale the upper limit of each class along the *X*-axis and frequencies along the *Y*-axis. To provide additional information, you can label the vertical axis on the right in terms of relative frequencies. In the Applewood Auto Group, the vertical axis on the left is labeled

from 0 to 180 and on the right from 0 to 100%. Note, as an example, that 50% on the right axis should be opposite 90 vehicles on the left axis and 100% on the right axis should be opposite 180 on the left axis.

To begin, the first plot is at  $x = 200$  and  $y = 0$ . None of the vehicles sold for a profit of less than \$200. The profit on 8 vehicles was less than \$600, so the next plot is at  $x = 600$  and  $y = 8$ . Continuing, the next plot is  $x = 1,000$  and  $y = 19$ . There were 19 vehicles that sold for a profit of less than \$1,000. The rest of the points are plotted and then the dots connected to form Chart 2–7.

We should point out that the shape of the distribution is the same if we use cumulative relative frequencies instead of the cumulative frequencies. The only difference is that the vertical axis is scaled in percentages. In the following charts, a percentage scale is added to the right side of the graphs to help answer questions about cumulative relative frequencies.



**CHART 2–7** Cumulative Frequency Polygon for Profit on Vehicles Sold Last Month at Applewood Auto Group

Using Chart 2–7 to find the amount of profit on 75% of the cars sold, draw a horizontal line from the 75% mark on the right-hand vertical axis over to the polygon, then drop down to the X-axis and read the amount of profit. The value on the X-axis is about \$2,300, so we estimate that 75% of the vehicles sold earned a profit of \$2,300 or less for the Applewood group.

To find the highest profit earned on 60 of the 180 vehicles, we use Chart 2–7 to locate the value of 60 on the left-hand vertical axis. Next, we draw a horizontal line from the value of 60 to the polygon and then drop down to the X-axis and read the profit. It is about \$1,600, so we estimate that 60 of the vehicles sold for a profit of less than \$1,600. We can also make estimates of the percentage of vehicles that sold for less than a particular amount. To explain, suppose we want to estimate the percentage of vehicles that sold for a profit of less than \$2,000. We begin by locating the value of \$2,000 on the X-axis, move vertically to the polygon, and then horizontally to the vertical axis on the right. The value is about 56%, so we conclude 56% of the vehicles sold for a profit of less than \$2,000.

## SELF-REVIEW 2-5



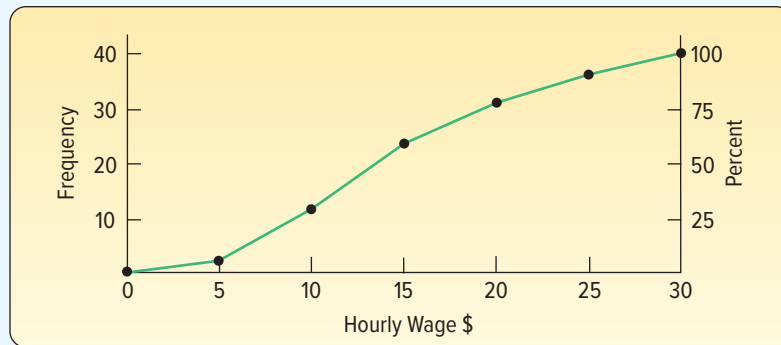
The hourly wages of the 15 employees of Matt's Tire and Auto Repair are organized into the following table.

| Hourly Wages    | Number of Employees |
|-----------------|---------------------|
| \$ 8 up to \$10 | 3                   |
| 10 up to 12     | 7                   |
| 12 up to 14     | 4                   |
| 14 up to 16     | 1                   |

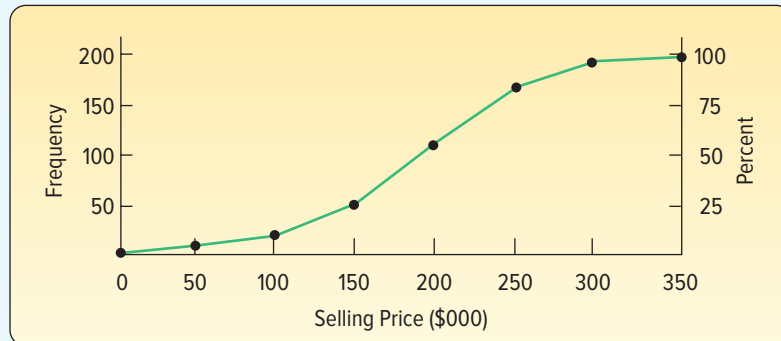
- What is the table called?
- Develop a cumulative frequency distribution and portray the distribution in a cumulative frequency polygon.
- On the basis of the cumulative frequency polygon, how many employees earn less than \$11 per hour?

## EXERCISES

19. The following cumulative frequency and the cumulative relative frequency polygon for the distribution of hourly wages of a sample of certified welders in the Atlanta, Georgia, area is shown in the graph.



- How many welders were studied?
  - What is the class interval?
  - About how many welders earn less than \$10.00 per hour?
  - About 75% of the welders make less than what amount?
  - Ten of the welders studied made less than what amount?
  - What percent of the welders make less than \$20.00 per hour?
20. The cumulative frequency and the cumulative relative frequency polygon for a distribution of selling prices (\$000) of houses sold in the Billings, Montana, area is shown in the graph.



- a. How many homes were studied?
  - b. What is the class interval?
  - c. One hundred homes sold for less than what amount?
  - d. About 75% of the homes sold for less than what amount?
  - e. Estimate the number of homes in the \$150,000 up to \$200,000 class.
  - f. About how many homes sold for less than \$225,000?
21. The frequency distribution representing the number of frequent flier miles accumulated by employees at Brumley Statistical Consulting Inc. is repeated from Exercise 17.

| Frequent Flier Miles<br>(000) | Frequency |
|-------------------------------|-----------|
| 0 up to 3                     | 5         |
| 3 up to 6                     | 12        |
| 6 up to 9                     | 23        |
| 9 up to 12                    | 8         |
| 12 up to 15                   | <u>2</u>  |
| Total                         | 50        |

- a. How many employees accumulated less than 3,000 miles?
  - b. Convert the frequency distribution to a cumulative frequency distribution.
  - c. Portray the cumulative distribution in the form of a cumulative frequency polygon.
  - d. Based on the cumulative relative frequencies, about 75% of the employees accumulated how many miles or less?
22. The frequency distribution of order lead time of the retailer from Exercise 18 is repeated below.

| Lead Time (days) | Frequency |
|------------------|-----------|
| 0 up to 5        | 6         |
| 5 up to 10       | 7         |
| 10 up to 15      | 12        |
| 15 up to 20      | 8         |
| 20 up to 25      | <u>7</u>  |
| Total            | 40        |

- a. How many orders were filled in less than 10 days? In less than 15 days?
- b. Convert the frequency distribution to cumulative frequency and cumulative relative frequency distributions.
- c. Develop a cumulative frequency polygon.
- d. About 60% of the orders were filled in less than how many days?

## CHAPTER SUMMARY

- I. A frequency table is a grouping of qualitative data into mutually exclusive and collectively exhaustive classes showing the number of observations in each class.
- II. A relative frequency table shows the fraction of the number of frequencies in each class.
- III. A bar chart is a graphic representation of a frequency table.
- IV. A pie chart shows the proportion each distinct class represents of the total number of observations.
- V. A frequency distribution is a grouping of data into mutually exclusive and collectively exhaustive classes showing the number of observations in each class.
  - A. The steps in constructing a frequency distribution are
    1. Decide on the number of classes.
    2. Determine the class interval.
    3. Set the individual class limits.
    4. Tally the raw data into classes and determine the frequency in each class.

- B. The class frequency is the number of observations in each class.
- C. The class interval is the difference between the limits of two consecutive classes.
- D. The class midpoint is halfway between the limits of consecutive classes.
- VI. A relative frequency distribution shows the percent of observations in each class.
- VII. There are several methods for graphically portraying a frequency distribution.
  - A. A histogram portrays the frequencies in the form of a rectangle or bar for each class. The height of the rectangles is proportional to the class frequencies.
  - B. A frequency polygon consists of line segments connecting the points formed by the intersection of the class midpoint and the class frequency.
  - C. A graph of a cumulative frequency distribution shows the number of observations less than a given value.
  - D. A graph of a cumulative relative frequency distribution shows the percent of observations less than a given value.

**CHAPTER EXERCISES**

- 23. Describe the similarities and differences of qualitative and quantitative variables. Be sure to include the following:
  - a. What level of measurement is required for each variable type?
  - b. Can both types be used to describe both samples and populations?
- 24. Describe the similarities and differences between a frequency table and a frequency distribution. Be sure to include which requires qualitative data and which requires quantitative data.
- 25. Alexandra Damonte will be building a new resort in Myrtle Beach, South Carolina. She must decide how to design the resort based on the type of activities that the resort will offer to its customers. A recent poll of 300 potential customers showed the following results about customers' preferences for planned resort activities:

|                                |     |
|--------------------------------|-----|
| Like planned activities        | 63  |
| Do not like planned activities | 135 |
| Not sure                       | 78  |
| No answer                      | 24  |

- a. What is the table called?
- b. Draw a bar chart to portray the survey results.
- c. Draw a pie chart for the survey results.
- d. If you are preparing to present the results to Ms. Damonte as part of a report, which graph would you prefer to show? Why?
- 26. **FILE** Speedy Swift is a package delivery service that serves the greater Atlanta, Georgia, metropolitan area. To maintain customer loyalty, one of Speedy Swift's performance objectives is on-time delivery. To monitor its performance, each delivery is measured on the following scale: early (package delivered before the promised time), on-time (package delivered within 15 minutes of the promised time), late (package delivered more than 15 minutes past the promised time), or lost (package never delivered). Speedy Swift's objective is to deliver 99% of all packages either early or on-time. Speedy collected the following data for last month's performance:

|         |         |         |         |         |         |         |         |         |         |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| On-time | On-time | Early   | Late    | On-time | On-time | On-time | On-time | Late    | On-time |
| Early   | On-time | On-time | Early   | On-time | On-time | On-time | On-time | On-time | On-time |
| Early   | On-time | Early   | On-time | On-time | On-time | Early   | On-time | On-time | On-time |
| Early   | On-time | On-time | Late    | Early   | Early   | On-time | On-time | On-time | Early   |
| On-time | Late    | Late    | On-time | On-time | On-time | On-time | On-time | On-time | On-time |
| On-time | Late    | Early   | On-time | Early   | On-time | Lost    | On-time | On-time | On-time |
| Early   | Early   | On-time | On-time | Late    | Early   | Lost    | On-time | On-time | On-time |
| On-time | On-time | Early   | On-time | Early   | On-time | Early   | On-time | Late    | On-time |
| On-time | Early   | On-time | On-time | On-time | Late    | On-time | Early   | On-time | On-time |
| On-time | On-time | On-time | On-time | On-time | Early   | Early   | On-time | On-time | On-time |

- a. What kind of variable is delivery performance? What scale is used to measure delivery performance?
  - b. Construct a frequency table for delivery performance for last month.
  - c. Construct a relative frequency table for delivery performance last month.
  - d. Construct a bar chart of the frequency table for delivery performance for last month.
  - e. Construct a pie chart of on-time delivery performance for last month.
  - f. Write a memo reporting the results of the analyses. Include your tables and graphs with written descriptions of what they show. Conclude with a general statement of last month's delivery performance as it relates to Speedy Swift's performance objectives.
27. A data set consists of 83 observations. How many classes would you recommend for a frequency distribution?
28. A data set consists of 145 observations that range from 56 to 490. What size class interval would you recommend?
29. **FILE** The following is the number of minutes to commute from home to work for a group of 25 automobile executives.

|    |    |    |    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 28 | 25 | 48 | 37 | 41 | 19 | 32 | 26 | 16 | 23 | 23 | 29 | 36 |
| 31 | 26 | 21 | 32 | 25 | 31 | 43 | 35 | 42 | 38 | 33 | 28 |    |

- a. How many classes would you recommend?
  - b. What class interval would you suggest?
  - c. What would you recommend as the lower limit of the first class?
  - d. Organize the data into a frequency distribution.
  - e. Comment on the shape of the frequency distribution.
30. **FILE** The following data give the weekly amounts spent on groceries for a sample of 45 households.

|       |       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| \$271 | \$363 | \$159 | \$ 76 | \$227 | \$337 | \$295 | \$319 | \$250 |
| 279   | 205   | 279   | 266   | 199   | 177   | 162   | 232   | 303   |
| 192   | 181   | 321   | 309   | 246   | 278   | 50    | 41    | 335   |
| 116   | 100   | 151   | 240   | 474   | 297   | 170   | 188   | 320   |
| 429   | 294   | 570   | 342   | 279   | 235   | 434   | 123   | 325   |

- a. How many classes would you recommend?
  - b. What class interval would you suggest?
  - c. What would you recommend as the lower limit of the first class?
  - d. Organize the data into a frequency distribution.
31. **FILE** A social scientist is studying the use of iPods by college students. A sample of 45 students revealed they played the following number of songs yesterday.

|   |   |   |    |   |   |   |   |   |   |   |   |   |   |    |
|---|---|---|----|---|---|---|---|---|---|---|---|---|---|----|
| 4 | 6 | 8 | 7  | 9 | 6 | 3 | 7 | 7 | 6 | 7 | 1 | 4 | 7 | 7  |
| 4 | 6 | 4 | 10 | 2 | 4 | 6 | 3 | 4 | 6 | 8 | 4 | 3 | 3 | 6  |
| 8 | 8 | 4 | 6  | 4 | 6 | 5 | 5 | 9 | 6 | 8 | 8 | 6 | 5 | 10 |

Organize the information into a frequency distribution.

- a. How many classes would you suggest?
  - b. What is the most suitable class interval?
  - c. What is the lower limit of the initial class?
  - d. Create the frequency distribution.
  - e. Describe the shape of the distribution.
32. **FILE** David Wise handles his own investment portfolio and has done so for many years. Listed below is the holding time (recorded to the nearest whole year) between purchase and sale for his collection of 36 stocks.

|   |    |   |    |    |   |   |   |    |   |    |    |    |   |    |   |   |    |   |   |
|---|----|---|----|----|---|---|---|----|---|----|----|----|---|----|---|---|----|---|---|
| 8 | 8  | 6 | 11 | 11 | 9 | 8 | 5 | 11 | 4 | 8  | 5  | 14 | 7 | 12 | 8 | 6 | 11 | 9 | 7 |
| 9 | 15 | 8 | 8  | 12 | 5 | 9 | 8 | 5  | 9 | 10 | 11 | 3  | 9 | 8  | 6 |   |    |   |   |

- a. How many classes would you propose?
- b. What class interval would you suggest?
- c. What quantity would you use for the lower limit of the initial class?

- d. Using your responses to parts (a), (b), and (c), create a frequency distribution.
- e. Describe the shape of the frequency distribution.

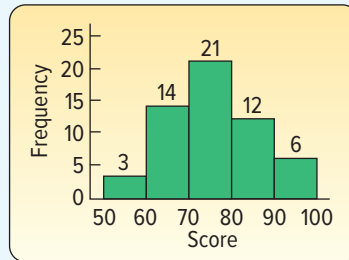
**33. FILE** You are exploring the music in your iTunes library. The total play counts over the past year for the 27 songs on your “smart playlist” are shown below. Make a frequency distribution of the counts and describe its shape. It is often claimed that a small fraction of a person’s songs will account for most of their total plays. Does this seem to be the case here?

|     |     |    |     |     |     |     |     |     |    |
|-----|-----|----|-----|-----|-----|-----|-----|-----|----|
| 128 | 56  | 54 | 91  | 190 | 23  | 160 | 298 | 445 | 50 |
| 578 | 494 | 37 | 677 | 18  | 74  | 70  | 868 | 108 | 71 |
| 466 | 23  | 84 | 38  | 26  | 814 | 17  |     |     |    |

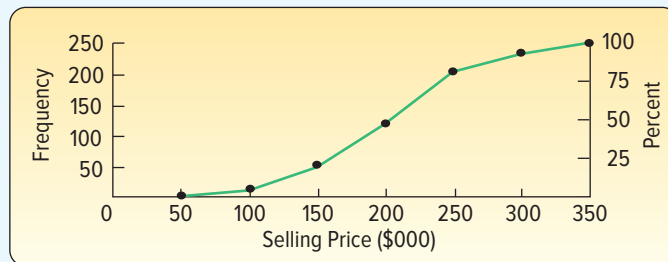
**34. FILE** The monthly issues of the *Journal of Finance* are available on the Internet. The table below shows the number of times an issue was downloaded over the last 33 months. Suppose that you wish to summarize the number of downloads with a frequency distribution.

|       |       |       |       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 312   | 2,753 | 2,595 | 6,057 | 7,624 | 6,624 | 6,362 | 6,575 | 7,760 | 7,085 | 7,272 |
| 5,967 | 5,256 | 6,160 | 6,238 | 6,709 | 7,193 | 5,631 | 6,490 | 6,682 | 7,829 | 7,091 |
| 6,871 | 6,230 | 7,253 | 5,507 | 5,676 | 6,974 | 6,915 | 4,999 | 5,689 | 6,143 | 7,086 |

- a. How many classes would you propose?
  - b. What class interval would you suggest?
  - c. What quantity would you use for the lower limit of the initial class?
  - d. Using your responses to parts (a), (b), and (c), create a frequency distribution.
  - e. Describe the shape of the frequency distribution.
- 35.** The following histogram shows the scores on the first exam for a statistics class.



- a. How many students took the exam?
  - b. What is the class interval?
  - c. What is the class midpoint for the first class?
  - d. How many students earned a score of less than 70?
- 36.** The following chart summarizes the selling price of homes sold last month in the Sarasota, Florida, area.



- a. What is the chart called?
- b. How many homes were sold during the last month?
- c. What is the class interval?
- d. About 75% of the houses sold for less than what amount?
- e. One hundred seventy-five of the homes sold for less than what amount?

37. **FILE** A chain of sport shops catering to beginning skiers, headquartered in Aspen, Colorado, plans to conduct a study of how much a beginning skier spends on his or her initial purchase of equipment and supplies. Based on these figures, it wants to explore the possibility of offering combinations, such as a pair of boots and a pair of skis, to induce customers to buy more. A sample of 44 cash register receipts revealed these initial purchases:

|       |       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| \$140 | \$ 82 | \$265 | \$168 | \$ 90 | \$114 | \$172 | \$230 | \$142 |
| 86    | 125   | 235   | 212   | 171   | 149   | 156   | 162   | 118   |
| 139   | 149   | 132   | 105   | 162   | 126   | 216   | 195   | 127   |
| 161   | 135   | 172   | 220   | 229   | 129   | 87    | 128   | 126   |
| 175   | 127   | 149   | 126   | 121   | 118   | 172   | 126   |       |

- Arrive at a suggested class interval.
  - Organize the data into a frequency distribution using a lower limit of \$70.
  - Interpret your findings.
38. **FILE** The numbers of outstanding shares for 24 publicly traded companies are listed in the following table.

| Company                  | Number of Outstanding Shares (millions) | Company                       | Number of Outstanding Shares (millions) |
|--------------------------|---|-------------------------------|---|
| Southwest Airlines       | 738                                     | Costco                        | 436                                     |
| FirstEnergy              | 418                                     | Home Depot                    | 1,495                                   |
| Harley-Davidson          | 226                                     | DTE Energy                    | 172                                     |
| Entergy                  | 178                                     | Dow Chemical                  | 1,199                                   |
| Chevron                  | 1,957                                   | Eastman Kodak                 | 272                                     |
| Pacific Gas and Electric | 430                                     | American Electric Power       | 485                                     |
| DuPont                   | 932                                     | ITT Corporation               | 93                                      |
| Westinghouse             | 22                                      | Ameren                        | 243                                     |
| Eversource               | 314                                     | Virginia Electric and Power   | 575                                     |
| Facebook                 | 1,067                                   | Public Service Electric & Gas | 506                                     |
| Google Inc.              | 64                                      | Consumers Energy              | 265                                     |
| Apple                    | 941                                     | Starbucks                     | 744                                     |

- Using the number of outstanding shares, summarize the companies with a frequency distribution.
  - Display the frequency distribution with a frequency polygon.
  - Create a cumulative frequency distribution of the outstanding shares.
  - Display the cumulative frequency distribution with a cumulative frequency polygon.
  - Based on the cumulative relative frequency distribution, 75% of the companies have less than what number of outstanding shares?
  - Write a brief analysis of this group of companies based on your statistical summaries of number of outstanding shares.
39. A recent survey showed that the typical American car owner spends \$2,950 per year on operating expenses. Below is a breakdown of the various expenditure items. Draw an appropriate chart to portray the data and summarize your findings in a brief report.

| Expenditure Item      | Amount  |
|-----------------------|---------|
| Fuel                  | \$ 603  |
| Interest on car loan  | 279     |
| Repairs               | 930     |
| Insurance and license | 646     |
| Depreciation          | 492     |
| Total                 | \$2,950 |



40. **FILE** Midland National Bank selected a sample of 40 student checking accounts. Below are their end-of-the-month balances.

|       |       |       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| \$404 | \$ 74 | \$234 | \$149 | \$279 | \$215 | \$123 | \$ 55 | \$ 43 | \$321 |
| 87    | 234   | 68    | 489   | 57    | 185   | 141   | 758   | 72    | 863   |
| 703   | 125   | 350   | 440   | 37    | 252   | 27    | 521   | 302   | 127   |
| 968   | 712   | 503   | 489   | 327   | 608   | 358   | 425   | 303   | 203   |

- Tally the data into a frequency distribution using \$100 as a class interval and \$0 as the starting point.
  - Draw a cumulative frequency polygon.
  - The bank considers any student with an ending balance of \$400 or more a “preferred customer.” Estimate the percentage of preferred customers.
  - The bank is also considering a service charge to the lowest 10% of the ending balances. What would you recommend as the cutoff point between those who have to pay a service charge and those who do not?
41. Residents of the state of South Carolina earned a total of \$69.5 billion in adjusted gross income. Seventy-three percent of the total was in wages and salaries; 11% in dividends, interest, and capital gains; 8% in IRAs and taxable pensions; 3% in business income pensions; 2% in Social Security; and the remaining 3% from other sources. Develop a pie chart depicting the breakdown of adjusted gross income. Write a paragraph summarizing the information.
42. **FILE** A recent study of home technologies reported the number of hours of personal computer usage per week for a sample of 60 persons. Excluded from the study were people who worked out of their home and used the computer as a part of their work.

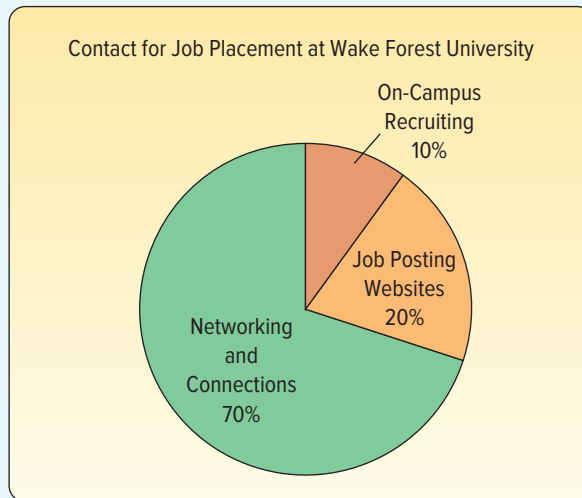
|     |     |     |      |     |     |     |     |     |     |
|-----|-----|-----|------|-----|-----|-----|-----|-----|-----|
| 9.3 | 5.3 | 6.3 | 8.8  | 6.5 | 0.6 | 5.2 | 6.6 | 9.3 | 4.3 |
| 6.3 | 2.1 | 2.7 | 0.4  | 3.7 | 3.3 | 1.1 | 2.7 | 6.7 | 6.5 |
| 4.3 | 9.7 | 7.7 | 5.2  | 1.7 | 8.5 | 4.2 | 5.5 | 5.1 | 5.6 |
| 5.4 | 4.8 | 2.1 | 10.1 | 1.3 | 5.6 | 2.4 | 2.4 | 4.7 | 1.7 |
| 2.0 | 6.7 | 1.1 | 6.7  | 2.2 | 2.6 | 9.8 | 6.4 | 4.9 | 5.2 |
| 4.5 | 9.3 | 7.9 | 4.6  | 4.3 | 4.5 | 9.2 | 8.5 | 6.0 | 8.1 |

- Organize the data into a frequency distribution. How many classes would you suggest? What value would you suggest for a class interval?
  - Draw a histogram. Describe your results.
43. **FILE** Merrill Lynch recently completed a study regarding the size of online investment portfolios (stocks, bonds, mutual funds, and certificates of deposit) for a sample of clients in the 40 up to 50 years old age group. Listed following is the value of all the investments in thousands of dollars for the 70 participants in the study.

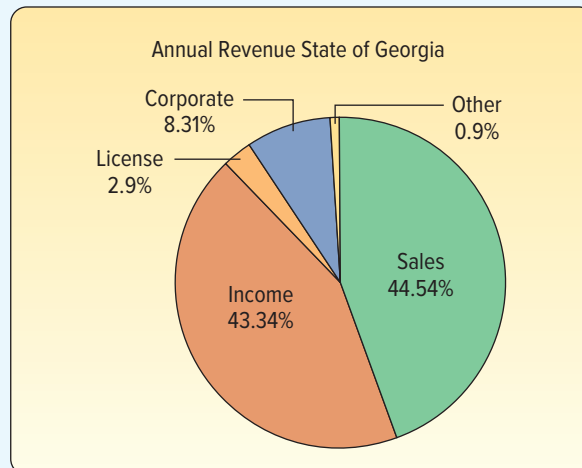
|         |        |         |        |         |         |          |         |
|---------|--------|---------|--------|---------|---------|----------|---------|
| \$669.9 | \$ 7.5 | \$ 77.2 | \$ 7.5 | \$125.7 | \$516.9 | \$ 219.9 | \$645.2 |
| 301.9   | 235.4  | 716.4   | 145.3  | 26.6    | 187.2   | 315.5    | 89.2    |
| 136.4   | 616.9  | 440.6   | 408.2  | 34.4    | 296.1   | 185.4    | 526.3   |
| 380.7   | 3.3    | 363.2   | 51.9   | 52.2    | 107.5   | 82.9     | 63.0    |
| 228.6   | 308.7  | 126.7   | 430.3  | 82.0    | 227.0   | 321.1    | 403.4   |
| 39.5    | 124.3  | 118.1   | 23.9   | 352.8   | 156.7   | 276.3    | 23.5    |
| 31.3    | 301.2  | 35.7    | 154.9  | 174.3   | 100.6   | 236.7    | 171.9   |
| 221.1   | 43.4   | 212.3   | 243.3  | 315.4   | 5.9     | 1,002.2  | 171.7   |
| 295.7   | 437.0  | 87.8    | 302.1  | 268.1   | 899.5   |          |         |

- Organize the data into a frequency distribution. How many classes would you suggest? What value would you suggest for a class interval?

- b. Draw a histogram. Financial experts suggest that this age group of people have at least five times their salary saved. As a benchmark, assume an investment portfolio of \$500,000 would support retirement in 10–15 years. In writing, summarize your results.
- 44. A total of 5.9% of the prime-time viewing audience watched shows on ABC, 7.6% watched shows on CBS, 5.5% on Fox, 6.0% on NBC, 2.0% on Warner Brothers, and 2.2% on UPN. A total of 70.8% of the audience watched shows on other cable networks, such as CNN and ESPN. You can find the latest information on TV viewing from the following website: <http://www.nielsen.com/us/en/top10s.html/>. Develop a pie chart or a bar chart to depict this information. Write a paragraph summarizing your findings.
- 45. Refer to the following chart:



- a. What is the name given to this type of chart?
- b. Suppose that 1,000 graduates will start a new job shortly after graduation. Estimate the number of graduates whose first contact for employment occurred through networking and other connections.
- c. Would it be reasonable to conclude that about 90% of job placements were made through networking, connections, and job posting websites? Cite evidence.
- 46. The following chart depicts the annual revenues, by type of tax, for the state of Georgia.



- a. What percentage of the state revenue is accounted for by sales tax and individual income tax?
  - b. Which category will generate more revenue: corporate taxes or license fees?
  - c. The total annual revenue for the state of Georgia is \$6.3 billion. Estimate the amount of revenue in billions of dollars for sales taxes and for individual taxes.
47. In 2016, the United States exported a total of \$266 billion worth of products to Canada. The five largest categories were:

| Product              | Amount |
|----------------------|--------|
| Vehicles             | \$48.1 |
| Machinery            | 40.0   |
| Electrical machinery | 23.9   |
| Mineral fuel and oil | 15.5   |
| Plastic              | 12.3   |

- a. Use a software package to develop a bar chart.
  - b. What percentage of the United States' total exports to Canada is represented by the two categories Machinery and Electrical Machinery?
  - c. What percentage of the top five exported products do Machinery and Electrical Machinery represent?
48. **FILE** In the United States, the industrial revolution of the early 20th century changed farming by making it more efficient. For example, in 1910, U.S. farms used 24.2 million horses and mules and only about 1,000 tractors. By 1960, 4.6 million tractors were used and only 3.2 million horses and mules. An outcome of making farming more efficient is the reduction of the number of farms from over 6 million in 1920 to about 2.2 million farms today. Listed below is the number of farms, in thousands, for each of the 50 states. Summarize the data and write a paragraph that describes your findings.

|    |    |    |     |    |    |    |    |    |    |
|----|----|----|-----|----|----|----|----|----|----|
| 50 | 12 | 5  | 28  | 59 | 19 | 35 | 22 | 80 | 5  |
| 8  | 48 | 3  | 75  | 25 | 77 | 46 | 68 | 10 | 69 |
| 77 | 25 | 13 | 20  | 35 | 6  | 52 | 61 | 36 | 38 |
| 88 | 1  | 75 | 246 | 59 | 50 | 44 | 98 | 74 | 2  |
| 32 | 42 | 7  | 31  | 28 | 9  | 8  | 44 | 25 | 37 |

49. One of the most popular candies in the United States is M&M's, produced by the Mars Company. In the beginning M&M's were all brown. Now they are produced in red, green, blue, orange, brown, and yellow. Recently, the purchase of a 14-ounce bag of M&M's Plain had 444 candies with the following breakdown by color: 130 brown, 98 yellow, 96 red, 35 orange, 52 blue, and 33 green. Develop a chart depicting this information and write a paragraph summarizing the results.
50. **FILE** The number of families who used the Minneapolis YWCA day care service was recorded during a 30-day period. The results are as follows:

|    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|
| 31 | 49 | 19 | 62 | 24 | 45 | 23 | 51 | 55 | 60 |
| 40 | 35 | 54 | 26 | 57 | 37 | 43 | 65 | 18 | 41 |
| 50 | 56 | 4  | 54 | 39 | 52 | 35 | 51 | 63 | 42 |

- a. Construct a cumulative frequency distribution.
- b. Sketch a graph of the cumulative frequency polygon.
- c. How many days saw fewer than 30 families utilize the day care center?
- d. Based on cumulative relative frequencies, how busy were the highest 80% of the days?

## DATA ANALYTICS

(The data for these exercises are available in Connect.)

51. **FILE** Refer to the North Valley Real Estate data, which report information on homes sold during the last year. For the variable *price*, select an appropriate class interval and organize the selling prices into a frequency distribution. Write a brief report summarizing your findings. Be sure to answer the following questions in your report.
- Around what values of price do the data tend to cluster?
  - Based on the frequency distribution, what is the typical selling price in the first class? What is the typical selling price in the last class?
  - Draw a cumulative relative frequency distribution. Using this distribution, fifty percent of the homes sold for what price or less? Estimate the lower price of the top ten percent of homes sold. About what percent of the homes sold for less than \$300,000?
  - Refer to the variable *bedrooms*. Draw a bar chart showing the number of homes sold with 2, 3, or 4 or more bedrooms. Write a description of the distribution.
52. **FILE** Refer to the Baseball 2016 data that report information on the 30 Major League Baseball teams for the 2016 season. Create a frequency distribution for the *Team Salary* variable and answer the following questions.
- What is the typical salary for a team? What is the range of the salaries?
  - Comment on the shape of the distribution. Does it appear that any of the teams have a salary that is out of line with the others?
  - Draw a cumulative relative frequency distribution of team salary. Using this distribution, forty percent of the teams have a salary of less than what amount? About how many teams have a total salary of more than \$220 million?
53. **FILE** Refer to the Lincolnville School District bus data. Select the variable referring to the number of *miles traveled since the last maintenance*, and then organize these data into a frequency distribution.
- What is a typical amount of miles traveled? What is the range?
  - Comment on the shape of the distribution. Are there any outliers in terms of miles driven?
  - Draw a cumulative relative frequency distribution. Forty percent of the buses were driven fewer than how many miles? How many buses were driven less than 10,500 miles?
  - Refer to the variables regarding the bus *manufacturer* and the bus *capacity*. Draw a pie chart of each variable and write a description of your results.

## PRACTICE TEST

## Part 1—Objective

- A grouping of *qualitative data* into mutually exclusive classes showing the number of observations in each class is known as a \_\_\_\_\_.
- A grouping of *quantitative data* into mutually exclusive classes showing the number of observations in each class is known as a \_\_\_\_\_.
- A graph in which the classes for qualitative data are reported on the horizontal axis and the class frequencies (proportional to the heights of the bars) on the vertical axis is called a \_\_\_\_\_.
- A circular chart that shows the proportion or percentage that each class represents of the total is called a \_\_\_\_\_.
- A graph in which the classes of a quantitative variable are marked on the horizontal axis and the class frequencies on the vertical axis is called a \_\_\_\_\_.
- A set of data included 70 observations. How many classes would you suggest to construct a frequency distribution?  
\_\_\_\_\_
- The distance between successive lower class limits is called the \_\_\_\_\_.
- The average of the respective class limits of two consecutive classes is the class \_\_\_\_\_.
- In a relative frequency distribution, the class frequencies are divided by the \_\_\_\_\_.
- A cumulative frequency polygon is created by line segments connecting the class \_\_\_\_\_ and the corresponding cumulative frequencies.

**Part 2—Problems**

1. Consider these data on the selling prices (\$000) of homes in the city of Warren, Pennsylvania, last year.

| Selling Price (\$000) | Frequency |
|-----------------------|-----------|
| \$120 up to \$150     | 4         |
| 150 up to 180         | 18        |
| 180 up to 210         | 30        |
| 210 up to 240         | 20        |
| 240 up to 270         | 17        |
| 270 up to 300         | 10        |
| 300 up to 330         | 6         |

- What is the class interval?
- How many homes were sold last year?
- How many homes sold for less than \$210,000?
- What is the relative frequency for the \$210 up to \$240 class?
- What is the midpoint of the \$150 up to \$180 class?
- What were the maximum and minimum selling prices?
- Construct a histogram of these data.
- Make a frequency polygon of these data.

# Describing Data:

## NUMERICAL MEASURES

# 3



©Andy Lyons/Getty Images

- ▲ **THE KENTUCKY DERBY** is held the first Saturday in May at Churchill Downs in Louisville, Kentucky. The racetrack is one and one-quarter miles. The table in Exercise 64 shows the winners since 1990, their margin of victory, the winning time, and the payoff on a \$2 bet. Determine the mean and median for the variables winning time and payoff on a \$2 bet. (See Exercise 64 and **LO3-1**.)

### LEARNING OBJECTIVES

*When you have completed this chapter, you will be able to:*

- LO3-1** Compute and interpret the mean, the median, and the mode.
- LO3-2** Compute a weighted mean.
- LO3-3** Compute and interpret the range, variance, and standard deviation.
- LO3-4** Explain and apply Chebyshev's theorem and the Empirical Rule.

## STATISTICS IN ACTION

Did you ever meet the “average” American man? Well, his name is Robert (that is the nominal level of measurement), and he is 31 years old (that is the ratio level), is 69.5 inches tall (again the ratio level of measurement), weighs 172 pounds, wears a size 9½ shoe, has a 34-inch waist, and wears a size 40 suit. In addition, the average man eats 4 pounds of potato chips, watches 1,456 hours of TV, and eats 26 pounds of bananas each year, and also sleeps 7.7 hours per night.

The average American woman is 5' 4" tall and weighs 140 pounds, while the average American model is 5' 11" tall and weighs 117 pounds. On any given day, almost half of the women in the United States are on a diet. Idolized in the 1950s, Marilyn Monroe would be considered overweight by today's standards. She fluctuated between a size 14 and a size 18 dress, and was a healthy and attractive woman.

## INTRODUCTION

Chapter 2 began our study of descriptive statistics. To summarize raw data into a meaningful form, we organized qualitative data into a frequency table and portrayed the results in a bar chart. In a similar fashion, we organized quantitative data into a frequency distribution and portrayed the results in a histogram. We also looked at other graphical techniques such as pie charts to portray qualitative data and frequency polygons to portray quantitative data.

This chapter is concerned with two numerical ways of describing quantitative variables, namely, **measures of location** and **measures of dispersion**. Measures of location are often referred to as averages. The purpose of a measure of location is to pinpoint the center of a distribution of data. An average is a measure of location that shows the central value of the data. Averages appear daily on TV, on various websites, in the newspaper, and in other journals. Here are some examples:

- The average U.S. home changes ownership every 11.8 years.
- An American receives an average of 568 pieces of mail per year.
- The average American home has more TV sets than people. There are 2.73 TV sets and 2.55 people in the typical home.
- The average American couple spends \$20,398 for their wedding, while their budget is 50% less. This does not include the cost of a honeymoon or engagement ring.
- The average price of a theater ticket in the United States is \$8.31, according to the National Association of Theater Owners.



©Andersen Ross/Getty Images RF

If we consider only measures of location in a set of data, or if we compare several sets of data using central values, we may draw an erroneous conclusion. In addition to measures of location, we should consider the **dispersion**—often called the *variation* or the *spread*—in the data. As an illustration, suppose the average annual income of executives for Internet-related companies is \$80,000, and the average income for executives in pharmaceutical firms is also \$80,000. If we looked only at the average incomes, we might wrongly conclude that the distributions of the two salaries are the same. However, we need to examine the dispersion or spread of the distributions of salary. A look at the salary ranges indicates that this conclusion of equal distributions is not correct. The salaries for the executives in the Internet firms range from \$70,000 to \$90,000, but salaries for the marketing executives in pharmaceuticals range from \$40,000 to \$120,000. Thus, we conclude that although the average salaries are the same for the two industries, there is much more spread or dispersion in salaries for the pharmaceutical executives. To describe the dispersion, we will consider the range, the variance, and the standard deviation.

## LO3-1

Compute and interpret the mean, the median, and the mode.

## MEASURES OF LOCATION

We begin by discussing measures of location. There is not just one measure of location; in fact, there are many. We will consider four: the arithmetic mean, the median, the mode, and the weighted mean. The arithmetic mean is the most widely used and widely reported measure of location. We study the mean as both a population parameter and a sample statistic.

## The Population Mean

Many studies involve all the values in a population. For example, there are 12 sales associates employed at the Reynolds Road Carpet Outlet. The mean amount of commission they earned last month was \$1,345. This is a population value because we considered the commission of *all* the sales associates. Other examples of a population mean would be:

- The mean closing price for Johnson & Johnson stock for the last 5 days is \$95.47.
- The mean number of overtime hours worked last week by the six welders employed by Butts Welding Inc. is 6.45 hours.
- Caryn Tirsch began a website last month devoted to organic gardening. The mean number of hits on her site for the 31 days in July was 84.36.

For raw data—that is, data that have not been grouped in a frequency distribution—the population mean is the sum of all the values in the population divided by the number of values in the population. To find the population mean, we use the following formula.

$$\text{Population mean} = \frac{\text{Sum of all the values in the population}}{\text{Number of values in the population}}$$

Instead of writing out in words the full directions for computing the population mean (or any other measure), it is more convenient to use the shorthand symbols of mathematics. The mean of the population using mathematical symbols is:

$$\text{POPULATION MEAN} \quad \mu = \frac{\sum x}{N} \quad (3-1)$$

where:

- $\mu$  represents the population mean. It is the Greek lowercase letter “mu.”
- $N$  is the number of values in the population.
- $x$  represents any particular value.
- $\Sigma$  is the Greek capital letter “sigma” and indicates the operation of adding.
- $\Sigma x$  is the sum of the  $x$  values in the population.

Any measurable characteristic of a population is called a **parameter**. The mean of a population is an example of a parameter.

**PARAMETER** A characteristic of a population.

### EXAMPLE

There are 42 exits on I-75 through the state of Kentucky. Listed below are the distances between exits (in miles).

|    |   |    |   |   |   |   |    |   |    |   |    |   |    |
|----|---|----|---|---|---|---|----|---|----|---|----|---|----|
| 11 | 4 | 10 | 4 | 9 | 3 | 8 | 10 | 3 | 14 | 1 | 10 | 3 | 5  |
| 2  | 2 | 5  | 6 | 1 | 2 | 2 | 3  | 7 | 1  | 3 | 7  | 8 | 10 |
| 1  | 4 | 7  | 5 | 2 | 2 | 5 | 1  | 1 | 3  | 3 | 1  | 2 | 1  |



Why is this information a population? What is the mean number of miles between exits?

### SOLUTION

This is a population because we are considering all the exits on I-75 in Kentucky. We add the distances between each of the 42 exits. The total distance is 192 miles. To find the arithmetic mean, we divide this total by 42. So the arithmetic mean is 4.57 miles, found by  $192/42$ . From formula (3–1):

$$\mu = \frac{\sum x}{N} = \frac{11 + 4 + 10 + \cdots + 1}{42} = \frac{192}{42} = 4.57$$

How do we interpret the value of 4.57? It is the typical number of miles between exits. Because we considered all the exits on I-75 in Kentucky, this value is a population parameter.



©Sheila Fitzgerald/Shutterstock

## The Sample Mean

As explained in Chapter 1, we often select a sample from the population to estimate a specific characteristic of the population. Smucker's quality assurance department needs to be assured that the amount of orange marmalade in the jar labeled as containing 12 ounces actually contains that amount. It would be very expensive and time-consuming to check the weight of each jar. Therefore, a sample of 20 jars is selected, the mean of the sample is determined, and that value is used to estimate the amount in each jar.

For raw data—that is, ungrouped data—the *mean is the sum of all the sampled values divided by the total number of sampled values*. To find the mean for a sample:

$$\text{Sample mean} = \frac{\text{Sum of all the values in the sample}}{\text{Number of values in the sample}}$$

The mean of a sample and the mean of a population are computed in the same way, but the shorthand notation used is different. The formula for the mean of a *sample* is:

**SAMPLE MEAN**

$$\bar{x} = \frac{\sum x}{n}$$

**(3–2)**

where:

- $\bar{x}$  represents the sample mean. It is read “x bar.”
- $n$  is the number of values in the sample.
- $x$  represents any particular value.
- $\Sigma$  is the Greek capital letter “sigma” and indicates the operation of adding.
- $\Sigma x$  is the sum of the  $x$  values in the sample.

The mean of a sample, or any other measure based on sample data, is called a **statistic**. If the mean weight of a sample of 10 jars of Smucker's orange marmalade is 11.5 ounces, this is an example of a statistic.

**STATISTIC** A characteristic of a sample.

### EXAMPLE

Verizon is studying the number of monthly minutes used by clients in a particular cell phone rate plan. A random sample of 12 clients showed the following number of minutes used last month.

|    |     |    |     |     |     |
|----|-----|----|-----|-----|-----|
| 90 | 77  | 94 | 89  | 119 | 112 |
| 91 | 110 | 92 | 100 | 113 | 83  |

What is the arithmetic mean number of minutes used last month?

### SOLUTION

Using formula (3–2), the sample mean is:

$$\begin{aligned}\text{Sample mean} &= \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}} \\ \bar{x} &= \frac{\Sigma x}{n} = \frac{90 + 77 + \cdots + 83}{12} = \frac{1,170}{12} = 97.5\end{aligned}$$

The arithmetic mean number of minutes used last month by the sample of cell phone users is 97.5 minutes.

## Properties of the Arithmetic Mean

The arithmetic mean is a widely used measure of location. It has several important properties:

1. **To compute a mean, the data must be measured at the interval or ratio level.** Recall from Chapter 1 that ratio-level data include such data as ages, incomes, and weights.
2. **All the values are included in computing the mean.**
3. **The mean is unique.** That is, there is only one mean in a set of data. Later in the chapter, we will discover a measure of location that might appear twice, or more than twice, in a set of data.
4. **The sum of the deviations of each value from the mean is zero.** Expressed symbolically:

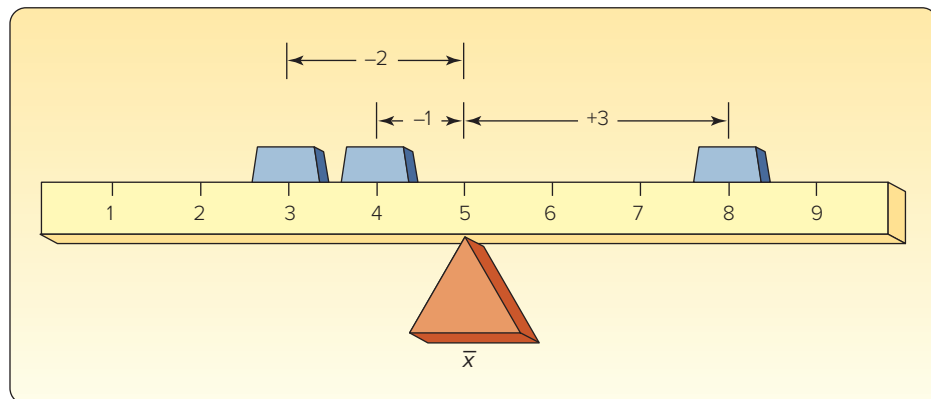
$$\Sigma(x - \bar{x}) = 0$$

As an example, the mean of 3, 8, and 4 is 5. Then:

$$\begin{aligned}\Sigma(x - \bar{x}) &= (3 - 5) + (8 - 5) + (4 - 5) \\ &= -2 + 3 - 1 \\ &= 0\end{aligned}$$

Thus, we can consider the mean as a balance point for a set of data. To illustrate, we have a long board with the numbers 1, 2, 3, . . . , 9 evenly spaced on it. Suppose three bars of equal weight were placed on the board at numbers 3, 4, and 8, and the balance point was set at 5, the mean of the three numbers. We would find that the board is

balanced perfectly! The deviations below the mean ( $-3$ ) are equal to the deviations above the mean ( $+3$ ). Shown schematically:



The mean does have a weakness. Recall that the mean uses the value of every item in a sample, or population, in its computation. If one or two of these values are either extremely large or extremely small compared to the majority of data, the mean might not be an appropriate average to represent the data. For example, suppose the annual incomes of a sample of financial planners at Merrill Lynch are \$62,900, \$61,600, \$62,500, \$60,800, and \$1,200,000. The mean income is \$289,560. Obviously, it is not representative of this group because all but one financial planner has an income in the \$60,000 to \$63,000 range. One income (\$1.2 million) is unduly affecting the mean.

## SELF-REVIEW 3-1



- The annual incomes of a sample of middle-management employees at Westinghouse are \$62,900, \$69,100, \$58,300, and \$76,800.
  - Give the formula for the sample mean.
  - Find the sample mean.
  - Is the mean you computed in (b) a statistic or a parameter? Why?
  - What is your best estimate of the population mean?
- The six students in Computer Science 411 are a population. Their final course grades are 92, 96, 61, 86, 79, and 84.
  - Give the formula for the population mean.
  - Compute the mean course grade.
  - Is the mean you computed in (b) a statistic or a parameter? Why?

## EXERCISES

*The answers to the odd-numbered exercises are in Appendix D.*

- Compute the mean of the following population values: 6, 3, 5, 7, 6.
- Compute the mean of the following population values: 7, 5, 7, 3, 7, 4.
- Compute the mean of the following sample values: 5, 9, 4, 10.
  - Show that  $\sum(x - \bar{x}) = 0$ .
- Compute the mean of the following sample values: 1.3, 7.0, 3.6, 4.1, 5.0.
  - Show that  $\sum(x - \bar{x}) = 0$ .
- Compute the mean of the following sample values: 16.25, 12.91, 14.58.
- Suppose you go to the grocery store and spend \$61.85 for the purchase of 14 items. What is the mean price per item?

For Exercises 7–10, (a) compute the arithmetic mean and (b) indicate whether it is a statistic or a parameter.

7. There are 10 salespeople employed by Midtown Ford. The number of new cars sold last month by the respective salespeople were: 15, 23, 4, 19, 18, 10, 10, 8, 28, 19.
8. A mail-order company counted the number of incoming calls per day to the company's toll-free number during the first 7 days in May: 14, 24, 19, 31, 36, 26, 17.
9. **FILE** The Cambridge Power and Light Company selected a random sample of 20 residential customers. Following are the amounts, to the nearest dollar, the customers were charged for electrical service last month:

|    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|
| 54 | 48 | 58 | 50 | 25 | 47 | 75 | 46 | 60 | 70 |
| 67 | 68 | 39 | 35 | 56 | 66 | 33 | 62 | 65 | 67 |

10. **FILE** A Human Resources manager at Metal Technologies studied the overtime hours of welders. A sample of 15 welders showed the following number of overtime hours worked last month.

|    |    |    |    |   |    |    |    |
|----|----|----|----|---|----|----|----|
| 13 | 13 | 12 | 15 | 7 | 15 | 5  | 12 |
| 6  | 7  | 12 | 10 | 9 | 13 | 12 |    |

11. AAA Heating and Air Conditioning completed 30 jobs last month with a mean revenue of \$5,430 per job. The president wants to know the total revenue for the month. Based on the limited information, can you compute the total revenue? What is it?
12. A large pharmaceutical company hires business administration graduates to sell its products. The company is growing rapidly and dedicates only 1 day of sales training for new salespeople. The company's goal for new salespeople is \$10,000 per month. The goal is based on the current mean sales for the entire company, which is \$10,000 per month. After reviewing the retention rates of new employees, the company finds that only 1 in 10 new employees stays longer than 3 months. Comment on using the current mean sales per month as a sales goal for new employees. Why do new employees leave the company?

## The Median

We have stressed that, for data containing one or two very large or very small values, the arithmetic mean may not be representative. The center for such data is better described by a measure of location called the **median**.

To illustrate the need for a measure of location other than the arithmetic mean, suppose you are seeking to buy a condominium in Palm Aire. Your real estate agent says that the typical price of the units currently available is \$110,000. Would you still want to look? If you had budgeted your maximum purchase price at \$75,000, you might think they are out of your price range. However, checking the prices of the individual units might change your mind. They are \$60,000, \$65,000, \$70,000, and \$80,000, and a superdeluxe penthouse costs \$275,000. The arithmetic mean price is \$110,000, as the real estate agent reported, but one price (\$275,000) is pulling the arithmetic mean upward, causing it to be an unrepresentative average. It does seem that a price around \$70,000 is a more typical or representative average, and it is. In cases such as this, the median provides a more valid measure of location.

**MEDIAN** The midpoint of the values after they have been ordered from the minimum to the maximum values.

The median price of the units available is \$70,000. To determine this, we order the prices from the minimum value (\$60,000) to the maximum value (\$275,000) and select the middle value (\$70,000). For the median, the data must be at least an ordinal level of measurement.

| Prices Ordered from<br>Minimum to Maximum |            | Prices Ordered from<br>Maximum to Minimum |
|---|------------|---|
| \$ 60,000                                 |            | \$275,000                                 |
| 65,000                                    |            | 80,000                                    |
| 70,000                                    | ← Median → | 70,000                                    |
| 80,000                                    |            | 65,000                                    |
| 275,000                                   |            | 60,000                                    |

Note that there is the same number of prices below the median of \$70,000 as above it. The median is, therefore, unaffected by extremely low or high prices. Had the highest price been \$90,000, or \$300,000, or even \$1 million, the median price would still be \$70,000. Likewise, had the lowest price been \$20,000 or \$50,000, the median price would still be \$70,000.

In the previous illustration, there are an *odd* number of observations (five). How is the median determined for an *even* number of observations? As before, the observations are ordered. Then by convention to obtain a unique value, we calculate the mean of the two middle observations. So for an even number of observations, the median may not be one of the given values.

### ▶ EXAMPLE

Facebook is a popular social networking website. Users can add friends, send them messages, and update their personal profiles to notify friends about themselves and their activities. A sample of 10 adults revealed they spent the following number of hours last month using Facebook.

3   5   7   5   9   1   3   9   17   10

Find the median number of hours.

### SOLUTION

Note that the number of adults sampled is even (10). The first step, as before, is to order the hours using Facebook from the minimum value to the maximum value. Then identify the two middle times. The arithmetic mean of the two middle observations gives us the median hours. Arranging the values from minimum to maximum:

1   3   3   5   5   7   9   9   10   17

The median is found by averaging the two middle values. The middle values are 5 hours and 7 hours, and the mean of these two values is 6. We conclude that the typical adult Facebook user spends 6 hours per month at the website. Notice that the median is not one of the values. Also, half of the times are below the median and half are above it.

The major properties of the median are:

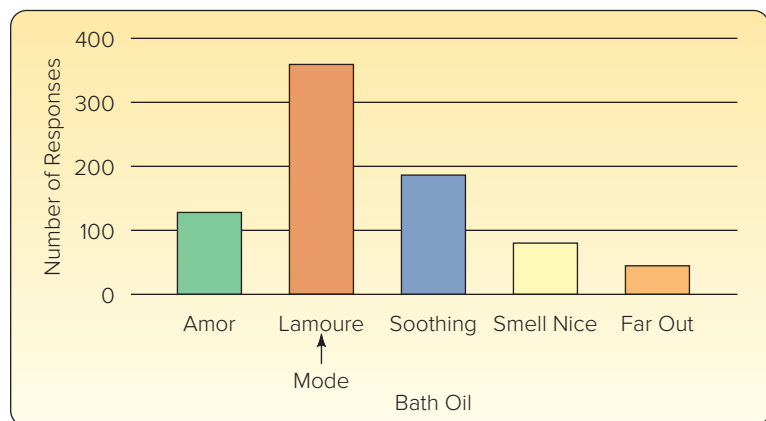
1. **It is not affected by extremely large or small values.** Therefore, the median is a valuable measure of location when such values do occur.
2. **It can be computed for ordinal-level data or higher.** Recall from Chapter 1 that ordinal-level data can be ranked from low to high.

## The Mode

The **mode** is another measure of location.

**MODE** The value of the observation that appears most frequently.

The mode is especially useful in summarizing nominal-level data. As an example of its use for nominal-level data, a company has developed five bath oils. The bar chart in Chart 3–1 shows the results of a marketing survey designed to find which bath oil consumers prefer. The largest number of respondents favored Lamoure, as evidenced by the highest bar. Thus, Lamoure is the mode.



**CHART 3–1** Number of Respondents Favoring Various Bath Oils

### EXAMPLE

Recall the data regarding the distance in miles between exits on I-75 in Kentucky. The information is repeated below.

|    |   |    |   |   |   |   |    |   |    |   |    |   |    |
|----|---|----|---|---|---|---|----|---|----|---|----|---|----|
| 11 | 4 | 10 | 4 | 9 | 3 | 8 | 10 | 3 | 14 | 1 | 10 | 3 | 5  |
| 2  | 2 | 5  | 6 | 1 | 2 | 2 | 3  | 7 | 1  | 3 | 7  | 8 | 10 |
| 1  | 4 | 7  | 5 | 2 | 2 | 5 | 1  | 1 | 3  | 3 | 1  | 2 | 1  |

What is the modal distance?

### SOLUTION

The first step is to organize the distances into a frequency table. This will help us determine the distance that occurs most frequently.

| Distance in Miles between Exits | Frequency |
|---------------------------------|-----------|
| 1                               | 8         |
| 2                               | 7         |
| 3                               | 7         |
| 4                               | 3         |
| 5                               | 4         |
| 6                               | 1         |
| 7                               | 3         |
| 8                               | 2         |
| 9                               | 1         |
| 10                              | 4         |
| 11                              | 1         |
| 14                              | 1         |
| Total                           | 42        |

The distance that occurs most often is 1 mile. This happens eight times—that is, there are eight exits that are 1 mile apart. So the modal distance between exits is 1 mile.

Which of the three measures of location (mean, median, or mode) best represents the central location of these data? Is the mode the best measure of location to represent the Kentucky data? No. The mode assumes only the nominal scale of measurement, and the variable miles is measured using the ratio scale. We calculated the mean to be 4.57 miles. See page 56. Is the mean the best measure of location to represent these data? Probably not. There are several cases in which the distance between exits is large. These values are affecting the mean, making it too large and not representative of the distances between exits. What about the median? The median distance is 3 miles. That is, half of the distances between exits are 3 miles or less. In this case, the median of 3 miles between exits is probably a more representative measure of the distance between exits.

In summary, we can determine the mode for all levels of data—nominal, ordinal, interval, and ratio. The mode also has the advantage of not being affected by extremely high or low values.

The mode does have disadvantages, however, that cause it to be used less frequently than the mean or median. For many sets of data, there is no mode because no value appears more than once. For example, there is no mode for this set of price data because every value occurs once: \$19, \$21, \$23, \$20, and \$18. Conversely, for some data sets there is more than one mode. Suppose the ages of the individuals in a stock investment club are 22, 26, 27, 27, 31, 35, and 35. Both the ages 27 and 35 are modes. Thus, this grouping of ages is referred to as *bimodal* (having two modes). One would question the use of two modes to represent the location of this set of age data.

## SELF-REVIEW 3-2



- A sample of single persons in Towson, Texas, receiving Social Security payments revealed these monthly benefits: \$852, \$598, \$580, \$1,374, \$960, \$878, and \$1,130.
  - What is the median monthly benefit?
  - How many observations are below the median? Above it?
- The number of work stoppages in the United States over the last 10 years are 22, 20, 21, 15, 5, 11, 19, 19, 15, and 11.
  - What is the median number of stoppages?
  - How many observations are below the median? Above it?
  - What is the modal number of work stoppages?

## EXERCISES

13. What would you report as the modal value for a set of observations if there were a total of:
- 10 observations and no two values were the same?
  - 6 observations and they were all the same?
  - 6 observations and the values were 1, 2, 3, 3, 4, and 4?

For Exercises 14–16, determine the (a) mean, (b) median, and (c) mode.

14. The following is the number of oil changes for the last 7 days at the Jiffy Lube located at the corner of Elm Street and Pennsylvania Avenue.

41 15 39 54 31 15 33

15. The following is the percent change in net income from last year to this year for a sample of 12 construction companies in Denver.

5 1 -10 -6 5 12 7 8 6 5 -1 11

16. The following are the ages of the 10 people in the Java Coffee Shop at the Southwyck Shopping Mall at 10 a.m.

21 41 20 23 24 33 37 42 23 29

17. **FILE** Several indicators of long-term economic growth in the United States and their annual percent change are listed below.

| Economic Indicator     | Percent Change | Economic Indicator           | Percent Change |
|------------------------|----------------|------------------------------|----------------|
| Inflation              | 4.5%           | Real GNP                     | 2.9%           |
| Exports                | 4.7            | Investment (residential)     | 3.6            |
| Imports                | 2.3            | Investment (nonresidential)  | 2.1            |
| Real disposable income | 2.9            | Productivity (total)         | 1.4            |
| Consumption            | 2.7            | Productivity (manufacturing) | 5.2            |

- What is the median percent change?
  - What is the modal percent change?
18. **FILE** Sally Reynolds sells real estate along the coastal area of Northern California. Below are her total annual commissions between 2007 and 2017. Find the mean, median, and mode of the commissions she earned for the 11 years.

| Year | Amount (thousands) |
|------|--------------------|
| 2007 | 292.16             |
| 2008 | 233.80             |
| 2009 | 206.97             |
| 2010 | 202.67             |
| 2011 | 164.69             |
| 2012 | 206.53             |
| 2013 | 237.51             |
| 2014 | 225.57             |
| 2015 | 255.33             |
| 2016 | 248.14             |
| 2017 | 269.11             |

19. **FILE** The accounting firm of Rowatti and Koppel specializes in income tax returns for self-employed professionals, such as physicians, dentists, architects, and lawyers. The firm employs 11 accountants who prepare the returns. For last year, the number of returns prepared by each accountant was:

58 75 31 58 46 65 60 71 45 58 80



Find the mean, median, and mode for the number of returns prepared by each accountant. If you could report only one, which measure of location would you recommend reporting?

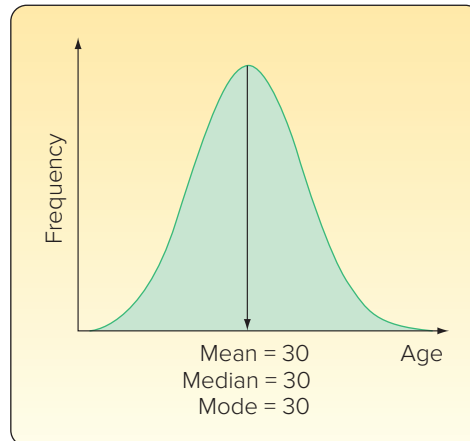
20. **FILE** The demand for the video games provided by Mid-Tech Video Games Inc. has exploded in the last several years. Hence, the owner needs to hire several new technical people to keep up with the demand. Mid-Tech gives each applicant a special test that Dr. McGraw, the designer of the test, believes is closely related to the ability to create video games. For the general population, the mean on this test is 100. Below are the scores on this test for the applicants.

95    105    120    81    90    115    99    100    130    10

The owner is interested in the overall quality of the job applicants based on this test. Compute the mean and the median scores for the 10 applicants. What would you report to the owner? Does it seem that the applicants are better than the general population?

## The Relative Positions of the Mean, Median, and Mode

Refer to the histogram of the variable age in Chart 3–2. It is a symmetric distribution, which is also mound-shaped. This distribution *has the same shape on either side of the center*. If the histogram were folded in half, the two halves would be identical. For any symmetric distribution, the mode, median, and mean are located at the center and are always equal. They are all equal to 30 years in Chart 3–2. We should point out that there are symmetric distributions that are not mound-shaped.

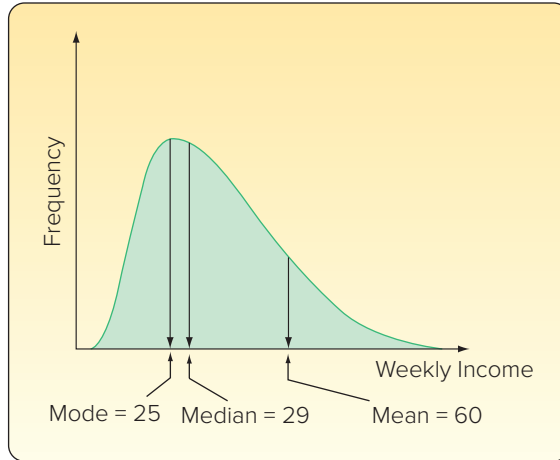


**CHART 3–2** A Symmetric Distribution

The number of years corresponding to the highest point of the curve is the *mode* (30 years). Because the distribution is symmetrical, the *median* corresponds to the point where the distribution is cut in half (30 years). Also, because the arithmetic mean is the balance point of a distribution (as shown in the Properties of the Arithmetic Mean section on page 58), and the distribution is symmetric, the arithmetic mean is 30. Logically, any of the three measures would be appropriate to represent the distribution's center.

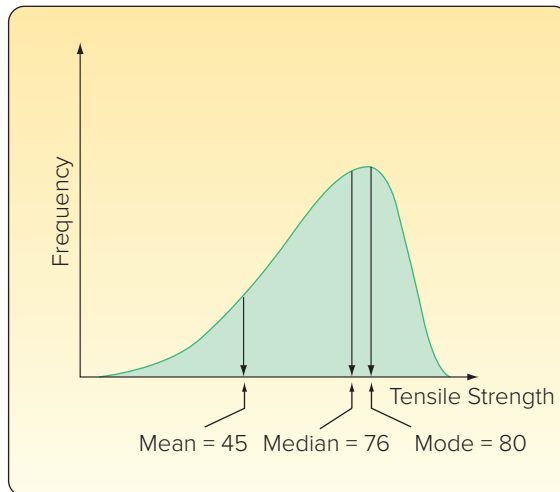
If a distribution is nonsymmetrical, or **skewed**, the relationship among the three measures changes. In a **positively skewed distribution**, such as the distribution of weekly income in Chart 3–3, the arithmetic mean is the largest of the three measures. Why? Because the mean is influenced by a few extremely high values. The median is generally the next largest measure in a positively skewed frequency distribution. The mode is the smallest of the three measures.

If the distribution is highly skewed, the mean would not be a good measure to use. The median and mode would be more representative.



**CHART 3-3** A Positively Skewed Distribution

Conversely, if a distribution is **negatively skewed**, such as the distribution of tensile strength in Chart 3-4, the mean is the lowest of the three measures. The mean is, of course, influenced by a few extremely low observations. The median is greater than the arithmetic mean, and the modal value is the largest of the three measures. Again, if the distribution is highly skewed, the mean should not be used to represent the data.



**CHART 3-4** A Negatively Skewed Distribution

**SELF-REVIEW 3-3**



The weekly sales from a sample of Hi-Tec electronic supply stores were organized into a frequency distribution. The mean of weekly sales was computed to be \$105,900, the median \$105,000, and the mode \$104,500.

- (a) Sketch the sales in the form of a smoothed frequency polygon. Note the location of the mean, median, and mode on the X-axis.
- (b) Is the distribution symmetrical, positively skewed, or negatively skewed? Explain.

**EXERCISES**

**21. FILE** The unemployment rate in the state of Alaska by month is given in the table below:

| Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 8.7 | 8.8 | 8.7 | 7.8 | 7.3 | 7.8 | 6.6 | 6.5 | 6.5 | 6.8 | 7.3 | 7.6 |

- a. What is the arithmetic mean of the Alaska unemployment rates?
  - b. Find the median and the mode for the unemployment rates.
  - c. Compute the arithmetic mean and median for just the winter (Dec–Mar) months. Is it much different?
22. **FILE** Big Orange Trucking is designing an information system for use in “in-cab” communications. It must summarize data from eight sites throughout a region to describe typical conditions. Compute an appropriate measure of central location for the variables wind direction, temperature, and pavement.

| City           | Wind Direction | Temperature | Pavement |
|----------------|----------------|-------------|----------|
| Anniston, AL   | West           | 89          | Dry      |
| Atlanta, GA    | Northwest      | 86          | Wet      |
| Augusta, GA    | Southwest      | 92          | Wet      |
| Birmingham, AL | South          | 91          | Dry      |
| Jackson, MS    | Southwest      | 92          | Dry      |
| Meridian, MS   | South          | 92          | Trace    |
| Monroe, LA     | Southwest      | 93          | Wet      |
| Tuscaloosa, AL | Southwest      | 93          | Trace    |

## Software Solution

We can use a statistical software package to find many measures of location.

### EXAMPLE

Table 2–4 on page 27 shows the profit on the sales of 180 vehicles at Applewood Auto Group. Determine the mean and the median selling price.

### SOLUTION

The mean, median, and modal amounts of profit are reported in the following output (highlighted in the screen shot). (Reminder: The instructions to create the output appear in the **Software Commands** in Appendix C.) There are 180 vehicles in the study, so using a calculator would be tedious and prone to error.

| APPLEWOOD AUTO GROUP |     |         |           |              |          |                    |           |   |
|----------------------|-----|---------|-----------|--------------|----------|--------------------|-----------|---|
|                      | A   | B       | C         | D            | E        | F                  | G         | H |
| 1                    | Age | Profit  | Location  | Vehicle-Type | Previous |                    | Profit    |   |
| 2                    | 21  | \$1,387 | Tionesta  | Sedan        | 0        |                    |           |   |
| 3                    | 23  | \$1,754 | Sheffield | SUV          | 1        | Mean               | 1843.17   |   |
| 4                    | 24  | \$1,817 | Sheffield | Hybrid       | 1        | Standard Error     | 47.97     |   |
| 5                    | 25  | \$1,040 | Sheffield | Compact      | 0        | Median             | 1882.50   |   |
| 6                    | 26  | \$1,273 | Kane      | Sedan        | 1        | Mode               | 1915.00   |   |
| 7                    | 27  | \$1,529 | Sheffield | Sedan        | 1        | Standard Deviation | 643.63    |   |
| 8                    | 27  | \$3,082 | Kane      | Truck        | 0        | Sample Variance    | 414256.61 |   |
| 9                    | 28  | \$1,951 | Kane      | SUV          | 1        | Kurtosis           | -0.22     |   |
| 10                   | 28  | \$2,692 | Tionesta  | Compact      | 0        | Skewness           | -0.24     |   |
| 11                   | 29  | \$1,342 | Kane      | Sedan        | 2        | Range              | 2998      |   |
| 12                   | 29  | \$1,206 | Sheffield | Sedan        | 0        | Minimum            | 294       |   |
| 13                   | 30  | \$443   | Kane      | Sedan        | 3        | Maximum            | 3292      |   |
| 14                   | 30  | \$1,621 | Sheffield | Truck        | 1        | Sum                | 331770    |   |
| 15                   | 30  | \$754   | Olean     | Sedan        | 2        | Count              | 180       |   |

Source: Microsoft Excel

The mean profit is \$1,843.17 and the median is \$1,882.50. These two values are less than \$40 apart, so either value is reasonable. We can also see from the Excel output that there were 180 vehicles sold and their total profit was \$331,770.00. We will describe the meaning of standard error, standard deviation, and other measures reported on the output later in this chapter and in later chapters.

What can we conclude? The typical profit on a vehicle is about \$1,850. Management at Applewood might use this value for revenue projections. For example, if the dealership could increase the number of vehicles sold in a month from 180 to 200, this would result in an additional estimated \$37,000 of revenue, found by  $20(\$1,850)$ .

**LO3-2**

Compute a weighted mean.

**THE WEIGHTED MEAN**

The weighted mean is a convenient way to compute the arithmetic mean when there are several observations of the same value. To explain, suppose the nearby Wendy's Restaurant sold medium, large, and Biggie-sized soft drinks for \$1.84, \$2.07, and \$2.40, respectively. Of the last 10 drinks sold, 3 were medium, 4 were large, and 3 were Biggie-sized. To find the mean price of the last 10 drinks sold, we could use formula (3–2).

$$\bar{x} = \frac{\$1.84 + \$1.84 + \$1.84 + \$2.07 + \$2.07 + \$2.07 + \$2.07 + \$2.40 + \$2.40 + \$2.40}{10}$$

$$\bar{x} = \frac{\$21.00}{10} = \$2.10$$

The mean selling price of the last 10 drinks is \$2.10.

An easier way to find the mean selling price is to determine the weighted mean. That is, we multiply each observation by the number of times it occurs. We will refer to the weighted mean as  $\bar{x}_w$ . This is read “x bar sub w.”

$$\bar{x}_w = \frac{3(\$1.84) + 4(\$2.07) + 3(\$2.40)}{10} = \frac{\$21.00}{10} = \$2.10$$

In this case, the weights are frequency counts. However, any measure of importance could be used as a weight. In general, the weighted mean of a set of numbers designated  $x_1, x_2, x_3, \dots, x_n$  with the corresponding weights  $w_1, w_2, w_3, \dots, w_n$  is computed by:

**WEIGHTED MEAN**

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \dots + w_nx_n}{w_1 + w_2 + w_3 + \dots + w_n} \quad (3-3)$$

This may be shortened to:

$$\bar{x}_w = \frac{\Sigma(wx)}{\Sigma w}$$

Note that the denominator of a weighted mean is always the sum of the weights.

**EXAMPLE**

The Carter Construction Company pays its hourly employees \$16.50, \$19.00, or \$25.00 per hour. There are 26 hourly employees, 14 of whom are paid at the \$16.50 rate, 10 at the \$19.00 rate, and 2 at the \$25.00 rate. What is the mean hourly rate paid the 26 employees?

**SOLUTION**

To find the mean hourly rate, we multiply each of the hourly rates by the number of employees earning that rate. From formula (3–3), the mean hourly rate is

$$\bar{x}_w = \frac{14(\$16.50) + 10(\$19.00) + 2(\$25.00)}{14 + 10 + 2} = \frac{\$471.00}{26} = \$18.1154$$

The weighted mean hourly wage is rounded to \$18.12.

**SELF-REVIEW 3–4**

Springers sold 95 Antonelli men's suits for the regular price of \$400. For the spring sale, the suits were reduced to \$200 and 126 were sold. At the final clearance, the price was reduced to \$100 and the remaining 79 suits were sold.

- What was the weighted mean price of an Antonelli suit?
- Springers paid \$200 a suit for the 300 suits. Comment on the store's profit per suit if a salesperson receives a \$25 commission for each one sold.

**EXERCISES**

- In June, an investor purchased 300 shares of Oracle (an information technology company) stock at \$41 per share. In August, she purchased an additional 400 shares at \$39 per share. In November, she purchased an additional 400 shares at \$45 per share. What is the weighted mean price per share?
- The Bookstall Inc. is a specialty bookstore concentrating on used books sold via the Internet. Paperbacks are \$1.00 each, and hardcover books are \$3.50. Of the 50 books sold last Tuesday morning, 40 were paperback and the rest were hardcover. What was the weighted mean price of a book?
- The Loris Healthcare System employs 200 persons on the nursing staff. Fifty are nurse's aides, 50 are practical nurses, and 100 are registered nurses. Nurse's aides receive \$8 an hour, practical nurses \$15 an hour, and registered nurses \$24 an hour. What is the weighted mean hourly wage?
- Andrews and Associates specialize in corporate law. They charge \$100 an hour for researching a case, \$75 an hour for consultations, and \$200 an hour for writing a brief. Last week one of the associates spent 10 hours consulting with her client, 10 hours researching the case, and 20 hours writing the brief. What was the weighted mean hourly charge for her legal services?

**LO3-3**

Compute and interpret the range, variance, and standard deviation.

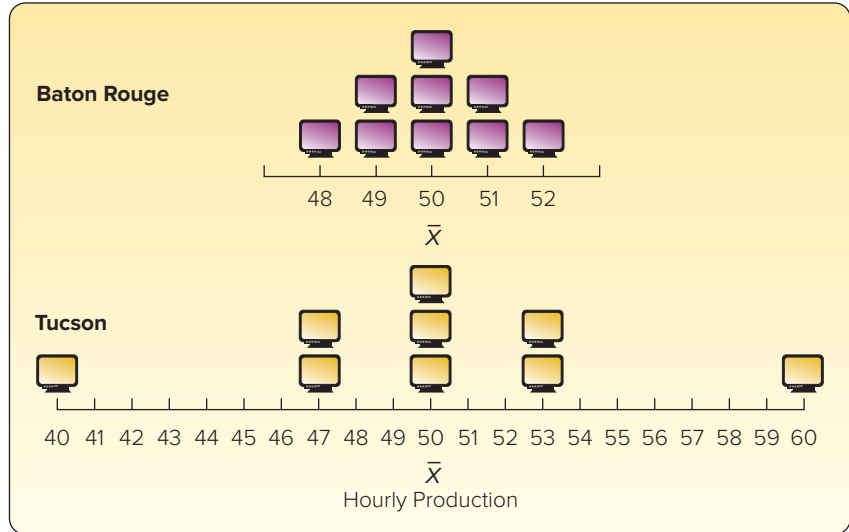
**WHY STUDY DISPERSION?**

A measure of location, such as the mean, median, or mode, only describes the center of the data. It is valuable from that standpoint, but it does not tell us anything about the spread of the data. For example, if your nature guide told you that the river ahead averaged 3 feet in depth, would you want to wade across on foot without additional information? Probably not. You would want to know something about the variation in the depth. Is the maximum depth of the river 3.25 feet and the minimum 2.75 feet? If that is the case, you would probably agree to cross. What if you learned the river depth ranged from 0.50 foot to 5.5 feet? Your decision would probably be not to cross. Before making a decision about crossing the river, you want information on both the typical depth and the dispersion in the depth of the river.

A reason to study dispersion is to compare the spread in two or more distributions. Suppose, for example, that the new Vision Quest LCD computer monitor is assembled

**STATISTICS IN ACTION**

The U.S. Postal Service has tried to become more “user friendly” in the last several years. A recent survey showed that customers were interested in more *consistency* in the time it takes to make a delivery. Under the old conditions, a local letter might take only one day to deliver, or it might take several. “Just tell me how many days ahead I need to mail the birthday card to Mom so it gets there on her birthday, not early, not late,” was a common complaint. The level of consistency is measured by the standard deviation of the delivery times.



**CHART 3-5** Hourly Production of Computer Monitors at the Baton Rouge and Tucson Plants

in Baton Rouge and also in Tucson. The arithmetic mean hourly output in both the Baton Rouge plant and the Tucson plant is 50. Based on the two means, you might conclude that the distributions of the hourly outputs are identical. Production records for 9 hours at the two plants, however, reveal that this conclusion is not correct (see Chart 3–5). Baton Rouge production varies from 48 to 52 assemblies per hour. Production at the Tucson plant is more erratic, ranging from 40 to 60 per hour. Therefore, the hourly output for Baton Rouge is clustered near the mean of 50; the hourly output for Tucson is more dispersed.

We will consider several measures of dispersion. The range is based on the maximum and minimum values in the data set; that is, only two values are considered. The variance and the standard deviation use all the values in a data set and are based on deviations from the arithmetic mean.

### Range

The simplest measure of dispersion is the **range**. It is the difference between the maximum and minimum values in a data set. Note that sometimes the range is interpreted as an interval. For example, the ages of high school students range between 12 and 20 years. In statistics, the range of ages would be 8 and calculated as follows:

**RANGE**                      Range = Maximum value – Minimum value                      **(3-4)**

The range is widely used in production management and control applications because it is very easy to calculate and understand.

**EXAMPLE**

Refer to Chart 3–5 above. Find the range in the number of computer monitors produced per hour for the Baton Rouge and the Tucson plants. Interpret the two ranges.

**SOLUTION**

The range of the hourly production of computer monitors at the Baton Rouge plant is 4, found by the difference between the maximum hourly production of 52 and the minimum of 48. The range in the hourly production for the Tucson plant is 20 computer monitors, found by  $60 - 40$ . We therefore conclude that there is less dispersion in the hourly production in the Baton Rouge plant than in the Tucson plant because the range of 4 computer monitors is less than a range of 20 computer monitors.

**Variance**

A limitation of the range is that it is based on only two values, the maximum and the minimum; it does not take into consideration all of the values. The **variance** does. It measures the mean amount by which the values in a population, or sample, vary from their mean. In terms of a definition:

**VARIANCE** The arithmetic mean of the squared deviations from the mean.

The following example illustrates how the variance is used to measure dispersion.

**EXAMPLE**

©Sorbis/Shutterstock

The chart below shows the number of cappuccinos sold at the Starbucks in the Orange County airport and the Ontario, California, airport between 4 and 5 p.m. for a sample of 5 days last month.

| California Airports |         |
|---------------------|---------|
| Orange County       | Ontario |
| 20                  | 20      |
| 40                  | 45      |
| 50                  | 50      |
| 60                  | 55      |
| 80                  | 80      |

Determine the mean, median, range, and variance for each location. Comment on the similarities and differences in these measures.

**SOLUTION**

The mean, median, and range for each of the airport locations are reported as part of an Excel spreadsheet.

|    | A      | B                   | C       |
|----|--------|---------------------|---------|
| 1  |        | California Airports |         |
| 2  |        | Orange County       | Ontario |
| 3  |        | 20                  | 20      |
| 4  |        | 40                  | 45      |
| 5  |        | 50                  | 50      |
| 6  |        | 60                  | 55      |
| 7  |        | 80                  | 80      |
| 8  |        |                     |         |
| 9  | Mean   | 50                  | 50      |
| 10 | Median | 50                  | 50      |
| 11 | Range  | 60                  | 60      |

Source: Microsoft

Notice that all three of the measures are exactly the same. Does this indicate that there is no difference in the two sets of data? We get a clearer picture if we calculate the variance. First, for Orange County:

| F   | G                 | H                 |
|---|-------------------|-------------------|
| Calculation of Variance for Orange County |                   |                   |
| Number Sold                               | Each Value - Mean | Squared Deviation |
| 20  | 20 - 50 = -30     | 900               |
| 40  | 40 - 50 = -10     | 100               |
| 50  | 50 - 50 = 0       | 0                 |
| 60  | 60 - 50 = 10      | 100               |
| 80  | 80 - 50 = 30      | 900               |
|   | Total             | 2000              |

Source: Microsoft

$$\text{Variance} = \frac{\sum(x - \mu)^2}{N} = \frac{(-30^2) + (-10^2) + 0^2 + 10^2 + 30^2}{5} = \frac{2,000}{5} = 400$$

The variance is 400. That is, the average squared deviation from the mean is 400.

The following shows the detail of determining the variance for the number of cappuccinos sold at the Ontario Airport.

| Calculation of Variance for Ontario |                   |                   |
|-------------------------------------|-------------------|-------------------|
| Number Sold                         | Each Value - Mean | Squared Deviation |
| 20                                  | 20 - 50 = -30     | 900               |
| 45                                  | 45 - 50 = -5      | 25                |
| 50                                  | 50 - 50 = 0       | 0                 |
| 55                                  | 55 - 50 = 5       | 25                |
| 80                                  | 80 - 50 = 30      | 900               |
|                                     | Total             | 1850              |

Source: Microsoft

$$\text{Variance} = \frac{\sum(x - \mu)^2}{N} = \frac{(-30^2) + (-5^2) + 0^2 + 5^2 + 30^2}{5} = \frac{1,850}{5} = 370$$

So the mean, median, and range of the cappuccinos sold are the same at the two airports, but the variances are different. The variance at Orange County is 400, but it is 370 at Ontario.

Let's interpret and compare the results of our measures for the two Starbucks airport locations. The mean and median of the two locations are exactly the same,



50 cappuccinos sold. These measures of location suggest the two distributions are the same. The range for both locations is also the same, 60. However, recall that the range provides limited information about the dispersion because it is based on only two values, the minimum and maximum.

The variances are not the same for the two airports. The variance is based on the differences between each observation and the arithmetic mean. It shows the closeness or clustering of the data relative to the mean or center of the distribution. Compare the variance for Orange County of 400 to the variance for Ontario of 370. Based on the variance, we conclude that the dispersion for the sales distribution of the Ontario Starbucks is more concentrated—that is, nearer the mean of 50—than for the Orange County location.

The variance has an important advantage over the range. It uses all the values in the computation. Recall that the range uses only the highest and the lowest values.

## SELF-REVIEW 3-5



The weights of containers being shipped to Ireland are (in thousands of pounds):

95    103    105    110    104    105    112    90

- What is the range of the weights?
- Compute the arithmetic mean weight.
- Compute the variance of the weights.

## EXERCISES

For Exercises 27–30, calculate the (a) range, (b) arithmetic mean, and (c) variance, and (d) interpret the statistics.

- FILE** During last weekend's sale, there were five customer service representatives on duty at the Electronic Super Store. The numbers of HDTVs these representatives sold were 5, 8, 4, 10, and 3.
- FILE** The Department of Statistics at Western State University offers eight sections of basic statistics. Following are the numbers of students enrolled in these sections: 34, 46, 52, 29, 41, 38, 36, and 28.
- FILE** Dave's Automatic Door installs automatic garage door openers. The following list indicates the number of minutes needed to install 10 door openers: 28, 32, 24, 46, 44, 40, 54, 38, 32, and 42.
- FILE** All eight companies in the aerospace industry were surveyed as to their return on investment last year. The results are: 10.6%, 12.6%, 14.8%, 18.2%, 12.0%, 14.8%, 12.2%, and 15.6%.
- FILE** Ten young adults living in California rated the taste of a newly developed sushi pizza, topped with tuna, rice, and kelp, on a scale of 1 to 50, with 1 indicating they did not like the taste and 50 that they did. The ratings were:

34    39    40    46    33    31    34    14    15    45

In a parallel study, 10 young adults in Iowa rated the taste of the same pizza. The ratings were:

28    25    35    16    25    29    24    26    17    20

- As a market researcher, compare the potential for sushi pizza in the two markets.
- FILE** The personnel files of all eight employees at the Pawnee location of Acme Carpet Cleaners Inc. revealed that during the last 6-month period they lost the following number of days due to illness:

2    0    6    3    10    4    1    2

All eight employees during the same period at the Chickpee location of Acme Carpet Cleaners revealed they lost the following number of days due to illness:

2 0 1 0 5 0 1 0

As the director of human resources, compare the two locations. What would you recommend?

## Population Variance

In the previous example, we developed the concept of variance as a measure of dispersion. Similar to the mean, we can calculate the variance of a population or the variance of a sample. The formula to compute the population variance is:

$$\text{POPULATION VARIANCE} \quad \sigma^2 = \frac{\sum(x - \mu)^2}{N} \quad (3-5)$$

where:

$\sigma^2$  is the population variance ( $\sigma$  is the lowercase Greek letter sigma). It is read as “sigma squared.”

$x$  is the value of a particular observation in the population.

$\mu$  is the arithmetic mean of the population.

$N$  is the number of observations in the population.

The process for computing the variance is implied by the formula.

1. Begin by finding the mean.
2. Find the difference between each observation and the mean, and square that difference.
3. Sum all the squared differences.
4. Divide the sum of the squared differences by the number of items in the population.

So the population variance is the mean of the squared difference between each value and the mean. For populations whose values are near the mean, the variance will be small. For populations whose values are dispersed from the mean, the population variance will be large.

The variance overcomes the weakness of the range by using all the values in the population, whereas the range uses only the maximum and minimum values. We overcome the issue where  $\sum(x - \mu) = 0$  by squaring the differences. Squaring the differences will always result in nonnegative values. The following is another example that illustrates the calculation and interpretation of the variance.

### EXAMPLE

The number of traffic citations issued last year by month in Beaufort County, South Carolina, is reported below.

| Citations by Month |          |       |       |     |      |      |        |           |         |          |          |
|--------------------|----------|-------|-------|-----|------|------|--------|-----------|---------|----------|----------|
| January            | February | March | April | May | June | July | August | September | October | November | December |
| 19                 | 17       | 22    | 18    | 28  | 34   | 45   | 39     | 38        | 44      | 34       | 10       |

Determine the population variance.

**SOLUTION**

Because we are studying all the citations for a year, the data comprise a population. To determine the population variance, we use formula (3–5). The table below details the calculations.

| Month     | Citations<br>( $x$ ) | $x - \mu$ | $(x - \mu)^2$ |
|-----------|----------------------|-----------|---------------|
| January   | 19                   | -10       | 100           |
| February  | 17                   | -12       | 144           |
| March     | 22                   | -7        | 49            |
| April     | 18                   | -11       | 121           |
| May       | 28                   | -1        | 1             |
| June      | 34                   | 5         | 25            |
| July      | 45                   | 16        | 256           |
| August    | 39                   | 10        | 100           |
| September | 38                   | 9         | 81            |
| October   | 44                   | 15        | 225           |
| November  | 34                   | 5         | 25            |
| December  | 10                   | -19       | 361           |
| Total     | 348                  | 0         | 1,488         |

1. We begin by determining the arithmetic mean of the population. The total number of citations issued for the year is 348, so the mean number issued per month is 29.

$$\mu = \frac{\sum x}{N} = \frac{19 + 17 + \cdots + 10}{12} = \frac{348}{12} = 29$$

2. Next we find the difference between each observation and the mean. This is shown in the third column of the table. The sum of the differences between the mean and the number of citations each month is 0. The principle is illustrated on page 58.
3. The next step is to square the difference for each month. That is shown in the fourth column of the table. All the squared differences will be positive. Note that squaring a negative value, or multiplying a negative value by itself, always results in a positive value.
4. The squared differences are totaled. The total of the fourth column is 1,488. That is the term  $\sum(x - \mu)^2$ .
5. Finally, we divide the squared differences by  $N$ , the number of observations in the population.

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N} = \frac{1,488}{12} = 124$$

So, the population variance for the number of citations is 124.

Like the range, the variance can be used to compare the dispersion in two or more sets of observations. For example, the variance for the number of citations issued in Beaufort County was just computed to be 124. If the variance in the number of citations issued in Marlboro County, South Carolina, is 342.9, we conclude that (1) there is less dispersion in the distribution of the number of citations issued in Beaufort County than in Marlboro County (because 124 is less than 342.9) and (2) the number of citations in Beaufort County is more closely clustered around the mean of 29 than the number of citations issued in Marlboro County. Thus the mean number of citations issued in Beaufort County is a more representative measure of location than the mean number of citations in Marlboro County.

## Population Standard Deviation

When we compute the variance, it is important to understand the unit of measure and what happens when the differences in the numerator are squared. That is, in the previous example, the number of monthly citations is the variable. When we calculate the variance, the unit of measure for the variance is citations squared. Using “squared citations” as a unit of measure is cumbersome.

There is a way out of this difficulty. By taking the square root of the population variance, we can transform it to the same unit of measurement used for the original data. The square root of 124 citations squared is 11.14 citations. The units are now simply citations. The square root of the population variance is the **population standard deviation**.

**POPULATION STANDARD DEVIATION**

$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}} \quad (3-6)$$

### SELF-REVIEW 3-6



The Philadelphia office of PricewaterhouseCoopers hired five accounting trainees this year. Their monthly starting salaries were \$3,536; \$3,173; \$3,448; \$3,121; and \$3,622.

- Compute the population mean.
- Compute the population variance.
- Compute the population standard deviation.
- The Pittsburgh office hired six trainees. Their mean monthly salary was \$3,550, and the standard deviation was \$250. Compare the two groups.

### EXERCISES

- Consider these five values a population: 8, 3, 7, 3, and 4.
  - Determine the mean of the population.
  - Determine the variance.
- Consider these six values a population: 13, 3, 8, 10, 8, and 6.
  - Determine the mean of the population.
  - Determine the variance.
- The annual report of Dennis Industries cited these primary earnings per common share for the past 5 years: \$2.68, \$1.03, \$2.26, \$4.30, and \$3.58. If we assume these are population values, what is:
  - The arithmetic mean primary earnings per share of common stock?
  - The variance?
- Referring to Exercise 35, the annual report of Dennis Industries also gave these returns on stockholder equity for the same 5-year period (in percent): 13.2, 5.0, 10.2, 17.5, and 12.9.
  - What is the arithmetic mean return?
  - What is the variance?
- Plywood Inc. reported these returns on stockholder equity for the past 5 years: 4.3, 4.9, 7.2, 6.7, and 11.6. Consider these as population values.
  - Compute the range, the arithmetic mean, the variance, and the standard deviation.
  - Compare the return on stockholder equity for Plywood Inc. with that for Dennis Industries cited in Exercise 36.
- The annual incomes of the five vice presidents of TMV Industries are \$125,000; \$128,000; \$122,000; \$133,000; and \$140,000. Consider this a population.
  - What is the range?
  - What is the arithmetic mean income?
  - What is the population variance? The standard deviation?
  - The annual incomes of officers of another firm similar to TMV Industries were also studied. The mean was \$129,000 and the standard deviation \$8,612. Compare the means and dispersions in the two firms.

**STATISTICS IN ACTION**

During the 2016 Major League Baseball season, DJ LeMahieu of the Colorado Rockies had the highest batting average at .348. Tony Gwynn hit .394 in the strike-shortened season of 1994, and Ted Williams hit .406 in 1941. No one has hit over .400 since 1941. The mean batting average has remained constant at about .260 for more than 100 years, but the standard deviation declined from .049 to .031. This indicates less dispersion in the batting averages today and helps explain the lack of any .400 hitters in recent times.

**Sample Variance and Standard Deviation**

The formula for the population mean is  $\mu = \Sigma x/N$ . We just changed the symbols for the sample mean; that is,  $\bar{x} = \Sigma x/n$ . Unfortunately, the conversion from the population variance to the sample variance is not as direct. It requires a change in the denominator. Instead of substituting  $n$  (number in the sample) for  $N$  (number in the population), the denominator is  $n - 1$ . Thus the formula for the **sample variance** is:

**SAMPLE VARIANCE**

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} \quad (3-7)$$

where:

- $s^2$  is the sample variance.
- $x$  is the value of each observation in the sample.
- $\bar{x}$  is the mean of the sample.
- $n$  is the number of observations in the sample.

Why is this change made in the denominator? Although the use of  $n$  is logical since  $\bar{x}$  is used to estimate  $\mu$ , it tends to underestimate the population variance,  $\sigma^2$ . The use of  $(n - 1)$  in the denominator provides the appropriate correction for this tendency. Because the primary use of sample statistics like  $s^2$  is to estimate population parameters like  $\sigma^2$ ,  $(n - 1)$  is preferred to  $n$  in defining the sample variance. We will also use this convention when computing the sample standard deviation.

**EXAMPLE**

The hourly wages for a sample of part-time employees at Pickett's Hardware Store are \$12, \$20, \$16, \$18, and \$19. What is the sample variance?

**SOLUTION**

The sample variance is computed by using formula (3-7).

$$\bar{x} = \frac{\Sigma x}{n} = \frac{\$85}{5} = \$17$$

| Hourly Wage<br>( $x$ ) | $x - \bar{x}$ | $(x - \bar{x})^2$ |
|------------------------|---------------|-------------------|
| \$12                   | -\$5          | 25                |
| 20                     | 3             | 9                 |
| 16                     | -1            | 1                 |
| 18                     | 1             | 1                 |
| 19                     | 2             | 4                 |
| <u>\$85</u>            | <u>0</u>      | <u>40</u>         |

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} = \frac{40}{5 - 1} = 10 \text{ in dollars squared}$$

The sample standard deviation is used as an estimator of the population standard deviation. As noted previously, the population standard deviation is the square root of the population variance. Likewise, the *sample standard deviation is the square root of the sample variance*. The sample standard deviation is determined by:

**SAMPLE STANDARD DEVIATION**  $s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$  **(3-8)**

**EXAMPLE**

The sample variance in the previous example involving hourly wages was computed to be 10. What is the sample standard deviation?

**SOLUTION**

The sample standard deviation is \$3.16, found by  $\sqrt{10}$ . Note again that the sample variance is in terms of dollars squared, but taking the square root of 10 gives us \$3.16, which is in the same units (dollars) as the original data.

**Software Solution**

On page 66, we used Excel to determine the mean, median, and mode of profit for the Applewood Auto Group data. You also will note that it lists the sample variance and sample standard deviation. Excel, like most other statistical software, assumes the data are from a sample.

| APPLEWOOD AUTO GROUP |     |         |           |              |          |                    |        |           |
|----------------------|-----|---------|-----------|--------------|----------|--------------------|--------|-----------|
|                      | A   | B       | C         | D            | E        | F                  | G      | H         |
| 1                    | Age | Profit  | Location  | Vehicle-Type | Previous |                    | Profit |           |
| 2                    | 21  | \$1,387 | Tionesta  | Sedan        | 0        |                    |        |           |
| 3                    | 23  | \$1,754 | Sheffield | SUV          | 1        | Mean               |        | 1843.17   |
| 4                    | 24  | \$1,817 | Sheffield | Hybrid       | 1        | Standard Error     |        | 47.97     |
| 5                    | 25  | \$1,040 | Sheffield | Compact      | 0        | Median             |        | 1882.50   |
| 6                    | 26  | \$1,273 | Kane      | Sedan        | 1        | Mode               |        | 1915.00   |
| 7                    | 27  | \$1,529 | Sheffield | Sedan        | 1        | Standard Deviation |        | 643.63    |
| 8                    | 27  | \$3,082 | Kane      | Truck        | 0        | Sample Variance    |        | 414256.61 |
| 9                    | 28  | \$1,951 | Kane      | SUV          | 1        | Kurtosis           |        | -0.22     |
| 10                   | 28  | \$2,692 | Tionesta  | Compact      | 0        | Skewness           |        | -0.24     |
| 11                   | 29  | \$1,342 | Kane      | Sedan        | 2        | Range              |        | 2998      |
| 12                   | 29  | \$1,206 | Sheffield | Sedan        | 0        | Minimum            |        | 294       |
| 13                   | 30  | \$443   | Kane      | Sedan        | 3        | Maximum            |        | 3292      |
| 14                   | 30  | \$1,621 | Sheffield | Truck        | 1        | Sum                |        | 331770    |
| 15                   | 30  | \$754   | Olean     | Sedan        | 2        | Count              |        | 180       |

Source: Microsoft Excel

**SELF-REVIEW 3-7**



The years of service for a sample of seven employees at a State Farm Insurance claims office in Cleveland, Ohio, are 4, 2, 5, 4, 5, 2, and 6. What is the sample variance? Compute the sample standard deviation.

## EXERCISES

For Exercises 39–44, do the following:

- a. Compute the sample variance.
  - b. Determine the sample standard deviation.
39. Consider these values a sample: 7, 2, 6, 2, and 3.
  40. The following five values are a sample: 11, 6, 10, 6, and 7.
  41. **FILE** Dave's Automatic Door, referred to in Exercise 29, installs automatic garage door openers. Based on a sample, following are the times, in minutes, required to install 10 door openers: 28, 32, 24, 46, 44, 40, 54, 38, 32, and 42.
  42. **FILE** The sample of eight companies in the aerospace industry, referred to in Exercise 30, was surveyed as to their return on investment last year. The results are 10.6, 12.6, 14.8, 18.2, 12.0, 14.8, 12.2, and 15.6.
  43. **FILE** The Houston, Texas, Motel Owner Association conducted a survey regarding weekday motel rates in the area. Listed below is the room rate for business-class guests for a sample of 10 motels.

\$101   \$97   \$103   \$110   \$78   \$87   \$101   \$80   \$106   \$88

44. **FILE** A consumer watchdog organization is concerned about credit card debt. A survey of 10 young adults with credit card debt of more than \$2,000 showed they paid an average of just over \$100 per month against their balances. Listed below are the amounts each young adult paid last month.

\$110   \$126   \$103   \$93   \$99   \$113   \$87   \$101   \$109   \$100

## LO3-4

Explain and apply Chebyshev's theorem and the Empirical Rule.

## INTERPRETATION AND USES OF THE STANDARD DEVIATION

The standard deviation is commonly used as a measure to compare the spread in two or more sets of observations. For example, the standard deviation of the biweekly amounts invested in the Dupree Paint Company profit-sharing plan is computed to be \$7.51. Suppose these employees are located in Georgia. If the standard deviation for a group of employees in Texas is \$10.47, and the means are about the same, it indicates that the amounts invested by the Georgia employees are not dispersed as much as those in Texas (because  $\$7.51 < \$10.47$ ). Since the amounts invested by the Georgia employees are clustered more closely about the mean, the mean for the Georgia employees is a more reliable measure than the mean for the Texas group.

## STATISTICS IN ACTION

Most colleges report the "average class size." This information can be misleading because average class size can be found in several ways. If we find the number of students *in each class* at a particular university, the result is the mean number of students per class. If we compile a list of the class sizes for each student and find the mean class size, we might find the mean to be quite different. One school found the mean number of students in each of its 747 classes to be 40. But when

(continued)

### Chebyshev's Theorem

We have stressed that a small standard deviation for a set of values indicates that these values are located close to the mean. Conversely, a large standard deviation reveals that the observations are widely scattered about the mean. The Russian mathematician P. L. Chebyshev (1821–1894) developed a theorem that allows us to determine the minimum proportion of the values that lie within a specified number of standard deviations of the mean. For example, according to **Chebyshev's theorem**, at least three out of every four, or 75%, of the values must lie between the mean plus two standard deviations and the mean minus two standard deviations. This relationship applies regardless of the shape of the distribution. Further, at least eight of nine values, or 88.9%, will lie between plus three standard deviations and minus three standard deviations of the mean. At least 24 of 25 values, or 96%, will lie between plus and minus five standard deviations of the mean.

Chebyshev's theorem states:

**CHEBYSHEV'S THEOREM** For any set of observations (sample or population), the proportion of the values that lie within  $k$  standard deviations of the mean is at least  $1 - 1/k^2$ , where  $k$  is any value greater than 1.

**EXAMPLE**

Dupree Paint Company employees contribute a mean of \$51.54 to the company's profit-sharing plan every two weeks. The standard deviation of biweekly contributions is \$7.51. At least what percent of the contributions lie within plus 3.5 standard deviations and minus 3.5 standard deviations of the mean, that is between \$25.26 and \$77.83?

**SOLUTION**

About 92%, found by

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(3.5)^2} = 1 - \frac{1}{12.25} = 0.92$$

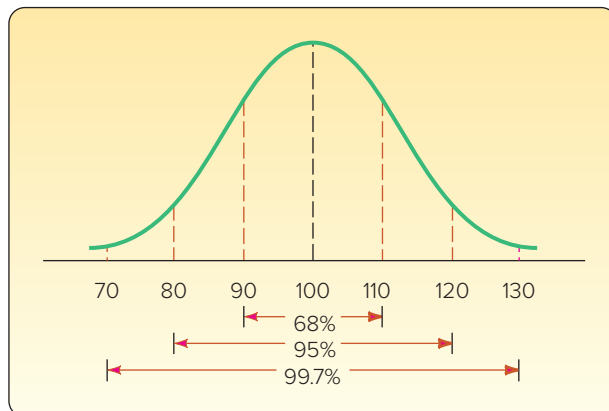
(continued from p. 78)  
it found the mean from a list of the class sizes of each student, it was 147. Why the disparity? Because there are few students in the small classes and a larger number of students in the larger classes, which has the effect of increasing the mean class size when it is calculated this way. A school could reduce this mean class size for each student by reducing the number of students in each class. That is, cut out the large freshman lecture classes.

**The Empirical Rule**

Chebyshev's theorem applies to any set of values; that is, the distribution of values can have any shape. However, for a symmetrical, bell-shaped distribution such as the one in Chart 3–6, we can be more precise in explaining the dispersion about the mean. These relationships involving the standard deviation and the mean are described by the **Empirical Rule**, sometimes called the **Normal Rule**.

**EMPIRICAL RULE** For a symmetrical, bell-shaped frequency distribution, approximately 68% of the observations lie within plus and minus one standard deviation of the mean; about 95% of the observations will lie within plus and minus two standard deviations of the mean; and practically all (99.7%) will lie within plus and minus three standard deviations of the mean.

These relationships are portrayed graphically in Chart 3–6 for a bell-shaped distribution with a mean of 100 and a standard deviation of 10.



**CHART 3–6** A Symmetrical, Bell-Shaped Curve Showing the Relationships between the Standard Deviation and the Percentage of Observations

Applying the Empirical Rule, if a distribution is symmetrical and bell-shaped, practically all of the observations lie between the mean plus and minus three standard deviations. Thus, if  $\bar{x} = 100$  and  $s = 10$ , practically all the observations lie between  $100 + 3(10)$  and  $100 - 3(10)$ , or 70 and 130. The estimated range is therefore 60, found by  $130 - 70$ .



Conversely, if we know that the range is 60 and the distribution is bell-shaped, we can approximate the standard deviation by dividing the range by 6. For this illustration:  $\text{range} \div 6 = 60 \div 6 = 10$ , the standard deviation.

### EXAMPLE

The monthly apartment rental rates near Crawford State University approximate a symmetrical, bell-shaped distribution. The sample mean is \$500; the standard deviation is \$20. Using the Empirical Rule, answer these questions:

1. About 68% of the monthly rentals are between what two amounts?
2. About 95% of the monthly rentals are between what two amounts?
3. Almost all of the monthly rentals are between what two amounts?

### SOLUTION

1. About 68% are between \$480 and \$520, found by  $\bar{x} \pm 1s = \$500 \pm 1(\$20)$ .
2. About 95% are between \$460 and \$540, found by  $\bar{x} \pm 2s = \$500 \pm 2(\$20)$ .
3. Almost all (99.7%) are between \$440 and \$560, found by  $\bar{x} \pm 3s = \$500 \pm 3(\$20)$ .

## SELF-REVIEW 3-8

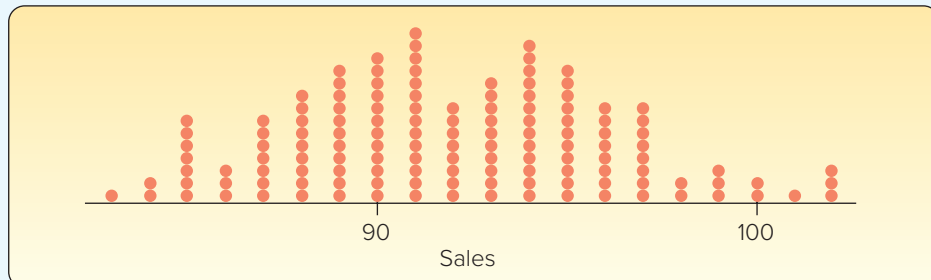


The Pitney Pipe Company is one of several domestic manufacturers of PVC pipe. The quality control department sampled 600 10-foot lengths. At a point 1 foot from the end of the pipe, they measured the outside diameter. The mean was 14.0 inches and the standard deviation was 0.1 inch.

- (a) If we do not know the shape of the distribution of outside pipe diameters, at least what percent of the observations will be between 13.85 inches and 14.15 inches?
- (b) If we assume that the distribution of diameters is symmetrical and bell-shaped, about 95% of the observations will be between what two values?

## EXERCISES

45. According to Chebyshev's theorem, at least what percent of any set of observations will be within 1.8 standard deviations of the mean?
46. The mean income of a group of sample observations is \$500; the standard deviation is \$40. According to Chebyshev's theorem, at least what percent of the incomes will lie between \$400 and \$600?
47. The distribution of the weights of a sample of 1,400 cargo containers is symmetric and bell-shaped. According to the Empirical Rule, what percent of the weights will lie:
  - a. Between  $\bar{x} - 2s$  and  $\bar{x} + 2s$ ?
  - b. Between  $\bar{x}$  and  $\bar{x} + 2s$ ? Above  $\bar{x} + 2s$ ?
48. The following graph portrays the distribution of the number of spicy chicken sandwiches sold at a nearby Wendy's for the last 141 days. The mean number of sandwiches sold per day is 91.9 and the standard deviation is 4.67.



If we use the Empirical Rule, sales will be between what two values on 68% of the days? Sales will be between what two values on 95% of the days?

## ETHICS AND REPORTING RESULTS

In Chapter 1, we discussed the ethical and unbiased reporting of statistical results. While you are learning about how to organize, summarize, and interpret data using statistics, it is also important to understand statistics so that you can be an intelligent consumer of information.

In this chapter, we learned how to compute numerical descriptive statistics. Specifically, we showed how to compute and interpret measures of location for a data set: the mean, median, and mode. We also discussed the advantages and disadvantages for each statistic. For example, if a real estate developer tells a client that the average home in a particular subdivision sold for \$150,000, we assume that \$150,000 is a representative selling price for all the homes. But suppose that the client also asks what the median sales price is, and the median is \$60,000. Why was the developer only reporting the mean price? This information is extremely important to a person's decision making when buying a home. Knowing the advantages and disadvantages of the mean, median, and mode is important as we report statistics and as we use statistical information to make decisions.

We also learned how to compute measures of dispersion: range, variance, and standard deviation. Each of these statistics also has advantages and disadvantages. Remember that the range provides information about the overall spread of a distribution. However, it does not provide any information about how the data are clustered or concentrated around the center of the distribution. As we learn more about statistics, we need to remember that when we use statistics we must maintain an independent and principled point of view. Any statistical report requires objective and honest communication of the results.

### CHAPTER SUMMARY

I. A measure of location is a value used to describe the central tendency of a set of data.

A. The arithmetic mean is the most widely reported measure of location.

1. It is calculated by adding the values of the observations and dividing by the total number of observations.

a. The formula for the population mean of ungrouped or raw data is

$$\mu = \frac{\sum X}{N} \quad (3-1)$$

b. The formula for the sample mean is

$$\bar{x} = \frac{\sum X}{n} \quad (3-2)$$

2. The major characteristics of the arithmetic mean are:

- At least the interval scale of measurement is required.
- All the data values are used in the calculation.
- A set of data has only one mean. That is, it is unique.
- The sum of the deviations from the mean equals 0.

B. The median is the value in the middle of a set of ordered data.

1. To find the median, sort the observations from minimum to maximum and identify the middle value.

2. The major characteristics of the median are:

- At least the ordinal scale of measurement is required.
- It is not influenced by extreme values.
- Fifty percent of the observations are larger than the median.
- It is unique to a set of data.

C. The mode is the value that occurs most often in a set of data.

- The mode can be found for nominal-level data.
- A set of data can have more than one mode.

- D. The weighted mean is found by multiplying each observation by its corresponding weight.
1. The formula for determining the weighted mean is

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n}{w_1 + w_2 + w_3 + \cdots + w_n} \quad (3-3)$$

- II. The dispersion is the variation or spread in a set of data.

- A. The range is the difference between the maximum and minimum values in a set of data.

1. The formula for the range is

$$\text{Range} = \text{Maximum value} - \text{Minimum value} \quad (3-4)$$

2. The major characteristics of the range are:

- a. Only two values are used in its calculation.
- b. It is influenced by extreme values.
- c. It is easy to compute and to understand.

- B. The variance is the mean of the squared deviations from the arithmetic mean.

1. The formula for the population variance is

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N} \quad (3-5)$$

2. The formula for the sample variance is

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} \quad (3-7)$$

3. The major characteristics of the variance are:

- a. All observations are used in the calculation.
- b. The units are somewhat difficult to work with; they are the original units squared.

- C. The standard deviation is the square root of the variance.

1. The major characteristics of the standard deviation are:

- a. It is in the same units as the original data.
- b. It is the square root of the average squared distance from the mean.
- c. It cannot be negative.
- d. It is the most widely reported measure of dispersion.

2. The formula for the sample standard deviation is

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} \quad (3-8)$$

- III. We use the standard deviation to describe a frequency distribution by applying Chebyshev's theorem or the Empirical Rule.

- A. Chebyshev's theorem states that regardless of the shape of the distribution, at least  $1 - 1/k^2$  of the observations will be within  $k$  standard deviations of the mean, where  $k$  is greater than 1.

- B. The Empirical Rule states that for a bell-shaped distribution about 68% of the values will be within one standard deviation of the mean, 95% within two, and virtually all within three.

## PRONUNCIATION KEY

| SYMBOL      | MEANING                       | PRONUNCIATION        |
|-------------|-------------------------------|----------------------|
| $\mu$       | Population mean               | <i>mu</i>            |
| $\Sigma$    | Operation of adding           | <i>sigma</i>         |
| $\Sigma x$  | Adding a group of values      | <i>sigma x</i>       |
| $\bar{x}$   | Sample mean                   | <i>x bar</i>         |
| $\bar{x}_w$ | Weighted mean                 | <i>x bar sub w</i>   |
| $\sigma^2$  | Population variance           | <i>sigma squared</i> |
| $\sigma$    | Population standard deviation | <i>sigma</i>         |

## CHAPTER EXERCISES

49. The accounting firm of Crawford and Associates has five senior partners. Yesterday the senior partners saw six, four, three, seven, and five clients, respectively.
- Compute the mean and median number of clients seen by the partners.
  - Is the mean a sample mean or a population mean?
  - Verify that  $\Sigma(x - \mu) = 0$ .
50. Owens Orchards sells apples in a large bag by weight. A sample of seven bags contained the following numbers of apples: 23, 19, 26, 17, 21, 24, 22.
- Compute the mean and median number of apples in a bag.
  - Verify that  $\Sigma(x - \bar{x}) = 0$ .
51. **FILE** A sample of households that subscribe to United Bell Phone Company for landline phone service revealed the following number of calls received per household last week. Determine the mean and the median number of calls received.

|    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|
| 52 | 43 | 30 | 38 | 30 | 42 | 12 | 46 | 39 | 37 |
| 34 | 46 | 32 | 18 | 41 | 5  |    |    |    |    |

52. **FILE** The Citizens Banking Company is studying the number of times the ATM located in a Loblaws Supermarket at the foot of Market Street is used per day. Following are the number of times the machine was used daily over each of the last 30 days. Determine the mean number of times the machine was used per day.

|    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|
| 83 | 64 | 84 | 76 | 84 | 54 | 75 | 59 | 70 | 61 |
| 63 | 80 | 84 | 73 | 68 | 52 | 65 | 90 | 52 | 77 |
| 95 | 36 | 78 | 61 | 59 | 84 | 95 | 47 | 87 | 60 |

53. **FILE** A recent study of the laundry habits of Americans included the time in minutes of the wash cycle. A sample of 40 observations follows. Determine the mean and the median of a typical wash cycle.

|    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|
| 35 | 37 | 28 | 37 | 33 | 38 | 37 | 32 | 28 | 29 |
| 39 | 33 | 32 | 37 | 33 | 35 | 36 | 44 | 36 | 34 |
| 40 | 38 | 46 | 39 | 37 | 39 | 34 | 39 | 31 | 33 |
| 37 | 35 | 39 | 38 | 37 | 32 | 43 | 31 | 31 | 35 |

54. **FILE** Trudy Green works for the True-Green Lawn Company. Her job is to solicit lawn care business via the telephone. Listed below is the number of appointments she made in each of the last 25 hours of calling. What is the arithmetic mean number of appointments she made per hour? What is the median number of appointments per hour? Write a brief report summarizing the findings.

|   |   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9 | 5 | 2 | 6 | 5 | 6 | 4 | 4 | 7 | 2 | 3 | 6 | 3 |
| 4 | 4 | 7 | 8 | 4 | 4 | 5 | 5 | 4 | 8 | 3 | 3 |   |

55. The Split-A-Rail Fence Company sells three types of fence to homeowners in suburban Seattle, Washington. Grade A costs \$5.00 per running foot to install, Grade B costs \$6.50 per running foot, and Grade C, the premium quality, costs \$8.00 per running foot. Yesterday, Split-A-Rail installed 270 feet of Grade A, 300 feet of Grade B, and 100 feet of Grade C. What was the mean cost per foot of fence installed?
56. Rolland Poust is a sophomore in the College of Business at Scandia Tech. Last semester he took courses in statistics and accounting, 3 hours each, and earned an A in both. He earned a B in a 5-hour history course and a B in a 2-hour history of jazz course. In addition, he took a 1-hour course dealing with the rules of basketball so he could get his license to officiate high school basketball games. He got an A in this course. What was his GPA for the semester? Assume that he receives 4 points for an A, 3 for a B, and so on. What measure of central tendency did you calculate? What method did you use?

57. The table below shows the percent of the labor force that is unemployed and the size of the labor force for three counties in northwest Ohio. Jon Elsas is the Regional Director of Economic Development. He must present a report to several companies that are considering locating in northwest Ohio. What would be an appropriate unemployment rate to show for the entire region?

| County | Percent Unemployed | Size of Workforce |
|--------|--------------------|-------------------|
| Wood   | 4.5                | 15,300            |
| Ottawa | 3.0                | 10,400            |
| Lucas  | 10.2               | 150,600           |

58. **FILE** The American Diabetes Association recommends a blood glucose reading of less than 130 for those with Type 2 diabetes. Blood glucose measures the amount of sugar in the blood. Below are the readings for February for a person recently diagnosed with Type 2 diabetes.

|     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 112 | 122 | 116 | 103 | 112 | 96  | 115 | 98  | 106 | 111 |
| 106 | 124 | 116 | 127 | 116 | 108 | 112 | 112 | 121 | 115 |
| 124 | 116 | 107 | 118 | 123 | 109 | 109 | 106 |     |     |

- What is the arithmetic mean glucose reading?
  - What is the median glucose reading?
  - What is the modal glucose reading?
59. The ages of a sample of Canadian tourists flying from Toronto to Hong Kong were 32, 21, 60, 47, 54, 17, 72, 55, 33, and 41.
- Compute the range.
  - Compute the standard deviation.
60. The weights (in pounds) of a sample of five boxes being sent by UPS are 12, 6, 7, 3, and 10.
- Compute the range.
  - Compute the standard deviation.
61. **FILE** The enrollments of the 13 public universities in the state of Ohio are listed below.

| College                        | Enrollment |
|--------------------------------|------------|
| University of Akron            | 26,106     |
| Bowling Green State University | 18,864     |
| Central State University       | 1,718      |
| University of Cincinnati       | 44,354     |
| Cleveland State University     | 17,194     |
| Kent State University          | 41,444     |
| Miami University               | 23,902     |
| Ohio State University          | 62,278     |
| Ohio University                | 36,493     |
| Shawnee State University       | 4,230      |
| University of Toledo           | 20,595     |
| Wright State University        | 17,460     |
| Youngstown State University    | 12,512     |

- Is this a sample or a population?
- What is the mean enrollment?
- What is the median enrollment?
- What is the range of the enrollments?
- Compute the standard deviation.

- 62. FILE** Creek Ratz is a very popular restaurant located along the coast of northern Florida. They serve a variety of steak and seafood dinners. During the summer beach season, they do not take reservations or accept “call ahead” seating. Management of the restaurant is concerned with the time a patron must wait before being seated for dinner. Listed below is the wait time, in minutes, for the 25 tables seated last Saturday night.

|    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|
| 28 | 39 | 23 | 67 | 37 | 28 | 56 | 40 | 28 | 50 |
| 51 | 45 | 44 | 65 | 61 | 27 | 24 | 61 | 34 | 44 |
| 64 | 25 | 24 | 27 | 29 |    |    |    |    |    |

- Explain why the times are a population.
  - Find the mean and median of the times.
  - Find the range and the standard deviation of the times.
- 63. FILE** A sample of 25 undergraduates reported the following dollar amounts of entertainment expenses last year:

|     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 684 | 710 | 688 | 711 | 722 | 698 | 723 | 743 | 738 | 722 | 696 | 721 | 685 |
| 763 | 681 | 731 | 736 | 771 | 693 | 701 | 737 | 717 | 752 | 710 | 697 |     |

- Find the mean, median, and mode of this information.
  - What are the range and standard deviation?
  - Use the Empirical Rule to establish an interval that includes about 95% of the observations.
- 64. FILE** The Kentucky Derby is held the first Saturday in May at Churchill Downs in Louisville, Kentucky. The racetrack is one and one-quarter miles. The following table shows the winners since 1990, their margin of victory, the winning time, and the payoff on a \$2 bet.

| Year | Winner            | Winning Margin (lengths) | Winning Time (minutes) | Payoff on a \$2 Win Bet |
|------|-------------------|--------------------------|------------------------|-------------------------|
| 1990 | Unbridled         | 3.5                      | 2.03333                | 10.80                   |
| 1991 | Strike the Gold   | 1.75                     | 2.05000                | 4.80                    |
| 1992 | Lil E. Tee        | 1                        | 2.05000                | 16.80                   |
| 1993 | Sea Hero          | 2.5                      | 2.04000                | 12.90                   |
| 1994 | Go For Gin        | 2                        | 2.06000                | 9.10                    |
| 1995 | Thunder Gulch     | 2.25                     | 2.02000                | 24.50                   |
| 1996 | Grindstone        | nose                     | 2.01667                | 5.90                    |
| 1997 | Silver Charm      | head                     | 2.04000                | 4.00                    |
| 1998 | Real Quiet        | 0.5                      | 2.03667                | 8.40                    |
| 1999 | Charismatic       | neck                     | 2.05333                | 31.30                   |
| 2000 | Fusaichi Pegasus  | 1.5                      | 2.02000                | 2.30                    |
| 2001 | Monarchos         | 4.75                     | 1.99950                | 10.50                   |
| 2002 | War Emblem        | 4                        | 2.01883                | 20.50                   |
| 2003 | Funny Cide        | 1.75                     | 2.01983                | 12.80                   |
| 2004 | Smarty Jones      | 2.75                     | 2.06767                | 4.10                    |
| 2005 | Giacomo           | 0.5                      | 2.04583                | 50.30                   |
| 2006 | Barbaro           | 6.5                      | 2.02267                | 6.10                    |
| 2007 | Street Sense      | 2.25                     | 2.03617                | 4.90                    |
| 2008 | Big Brown         | 4.75                     | 2.03033                | 6.80                    |
| 2009 | Mine That Bird    | 6.75                     | 2.04433                | 103.20                  |
| 2010 | Super Saver       | 2.50                     | 2.07417                | 18.00                   |
| 2011 | Animal Kingdom    | 2.75                     | 2.034                  | 43.80                   |
| 2012 | I'll Have Another | 1.5                      | 2.03050                | 32.60                   |
| 2013 | Orb               | 2.5                      | 2.04817                | 12.80                   |
| 2014 | California Chrome | 1.75                     | 2.0610                 | 7.00                    |
| 2015 | American Pharaoh  | 1.00                     | 2.05033                | 7.80                    |
| 2016 | Nyquist           | 1.25                     | 2.02183                | 6.60                    |

- a. Determine the mean and median for the variables winning time and payoff on a \$2 bet.
- b. Determine the range and standard deviation of the variables winning time and payoff on a \$2 bet.
- c. Refer to the variable winning margin. What is the level of measurement? What measure of location would be most appropriate?
65. **FILE** The manager of the local Walmart Supercenter is studying the number of items purchased by customers in the evening hours. Listed below is the number of items for a sample of 30 customers. Find the mean and the median of the number of items.

|    |   |    |    |    |    |    |    |    |    |
|----|---|----|----|----|----|----|----|----|----|
| 15 | 8 | 6  | 9  | 9  | 4  | 18 | 10 | 10 | 12 |
| 12 | 4 | 7  | 8  | 12 | 10 | 10 | 11 | 9  | 13 |
| 5  | 6 | 11 | 14 | 5  | 6  | 6  | 5  | 13 | 5  |

## DATA ANALYTICS

(The data for these exercises are available in Connect.)

66. **FILE** Refer to the North Valley Real Estate data and prepare a report on the sales prices of the homes. Be sure to answer the following questions in your report.
- a. Around what values of price do the data tend to cluster? What is the mean sales price? What is the median sales price? Is one measure more representative of the typical sales prices than the others?
- b. What is the range of sales prices? What is the standard deviation? About 95% of the sales prices are between what two values? Is the standard deviation a useful statistic for describing the dispersion of sales price?
- c. Repeat (a) and (b) using FICO score.
67. **FILE** Refer to the Baseball 2016 data, which report information on the 30 Major League Baseball teams for the 2016 season. Refer to the variable team salary. Prepare a report on the team salaries. Be sure to answer the following questions in your report.
- a. Around what values do the data tend to cluster? Specifically, what is the mean team salary? What is the median team salary? Is one measure more representative of the typical team salary than the others?
- b. What is the range of the team salaries? What is the standard deviation? About 95% of the salaries are between what two values?
68. **FILE** Refer to the Lincolnville School District bus data. Prepare a report on the maintenance cost for last month. Be sure to answer the following questions in your report.
- a. Around what values do the data tend to cluster? Specifically, what was the mean maintenance cost last month? What is the median cost? Is one measure more representative of the typical cost than the others?
- b. What is the range of maintenance costs? What is the standard deviation? About 95% of the maintenance costs are between what two values?

## PRACTICE TEST

### Part 1—Objective

1. An observable characteristic of a population is called a \_\_\_\_\_.
2. A measure, such as the mean, based on sample data is called a \_\_\_\_\_.
3. The sum of the differences between each value and the mean is always equal to \_\_\_\_\_.
4. The midpoint of a set of values after they have been ordered from the minimum to the maximum values is called the \_\_\_\_\_.
5. What percentage of the values in every data set is larger than the median? \_\_\_\_\_
6. The value of the observation that appears most frequently in a data set is called the \_\_\_\_\_.
7. The \_\_\_\_\_ is the difference between the maximum and minimum values in a data set.
8. The \_\_\_\_\_ is the arithmetic mean of the squared deviations from the mean.
9. The square of the standard deviation is the \_\_\_\_\_.
10. The standard deviation assumes a negative value when (all the values are negative, at least half the values are negative, or never—pick one) \_\_\_\_\_.

11. Which of the following is least affected by an outlier? (mean, median, or range—pick one) \_\_\_\_\_.
12. The \_\_\_\_\_ states that for any symmetrical, bell-shaped frequency distribution, approximately 68% of the observations will lie within plus and minus one standard deviation of the mean.

### Part 2—Problems

1. A sample of college students reported they owned the following number of CDs.

52 76 64 79 80 74 66 69

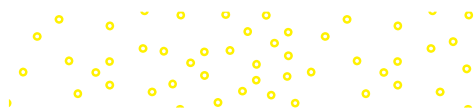
- a. What is the mean number of CDs owned?
  - b. What is the median number of CDs owned?
  - c. What is the range of the number of CDs owned?
  - d. What is the standard deviation of the number of CDs owned?
2. An investor purchased 200 shares of the Blair Company for \$36 each in July 2008, 300 shares at \$40 each in September 2008, and 500 shares at \$50 each in January 2009. What is the investor's weighted mean price per share?
  3. *The Wall Street Journal* regularly surveys a group of about 50 economists. Their forecasts for the change in the domestic gross national product (GNP) are normally distributed with a mean change of  $-0.88\%$ . That indicates a predicted decline in GNP of almost nine-tenths of a percent. If the standard deviation is  $1.41\%$ , use the Empirical Rule to estimate the range that includes 95% of the forecast changes in GNP.



# 4

# Describing Data:

## DISPLAYING AND EXPLORING DATA



©allstars/Shutterstock

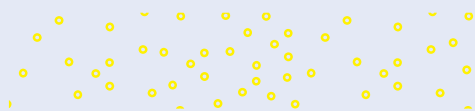
- ▲ **MCGIVERN JEWELERS** recently posted an advertisement on a social media site reporting the shape, size, price, and cut grade for 33 of its diamonds in stock. Develop a box plot of the variable price and comment on the result. (See Exercise 29 and **LO4-3**.)

## LEARNING OBJECTIVES

---

*When you have completed this chapter, you will be able to:*

- LO4-1** Construct and interpret a dot plot.
- LO4-2** Identify and compute measures of position.
- LO4-3** Construct and analyze a box plot.
- LO4-4** Compute and interpret the coefficient of skewness.
- LO4-5** Create and interpret a scatter diagram.
- LO4-6** Develop and explain a contingency table.



## INTRODUCTION

Chapter 2 began our study of descriptive statistics. In order to transform raw or ungrouped data into a meaningful form, we organize the data into a frequency distribution. We present the frequency distribution in graphic form as a histogram or a frequency polygon. This allows us to visualize where the data tend to cluster, the largest and the smallest values, and the general shape of the data.

In Chapter 3, we first computed several measures of location, such as the mean, median, and mode. These measures of location allow us to report a typical value in the set of observations. We also computed several measures of dispersion, such as the range, variance, and standard deviation. These measures of dispersion allow us to describe the variation or the spread in a set of observations.

We continue our study of descriptive statistics in this chapter. We study (1) dot plots, (2) percentiles, and (3) box plots. These charts and statistics give us additional insight into where the values are concentrated as well as the general shape of the data. Then we consider bivariate data. In bivariate data, we observe two variables for each individual or observation. Examples include the number of hours a student studied and the points earned on an examination; if a sampled product meets quality specifications and the shift on which it is manufactured; or the amount of electricity used in a month by a homeowner and the mean daily high temperature in the region for the month. These charts and graphs provide useful insights as we use business analytics to enhance our understanding of data.

### LO4-1

Construct and interpret a dot plot.

## DOT PLOTS

Recall for the Applewood Auto Group data, we summarized the profit earned on the 180 vehicles sold with a frequency distribution using eight classes. When we organized the data into the eight classes, we lost the exact value of the observations. A **dot plot**, on the other hand, groups the data as little as possible, and we do not lose the identity of an individual observation. To develop a dot plot, we display a dot for each observation along a horizontal number line indicating the possible values of the data. If there are identical observations or the observations are too close to be shown individually, the dots are “piled” on top of each other. This allows us to see the shape of the distribution, the value about which the data tend to cluster, and the largest and smallest observations. Dot plots are most useful for smaller data sets, whereas histograms tend to be most useful for large data sets. An example will show how to construct and interpret dot plots.

**DOT PLOT** A dot plot summarizes the distribution of one variable by stacking dots at points on a number line that shows the values of the variable. A dot plot shows all values.

### EXAMPLE

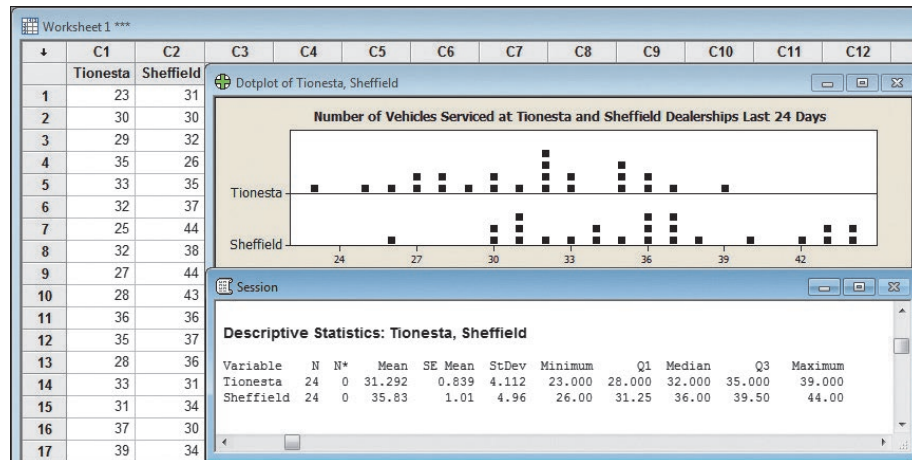
The service departments at Tionesta Ford Lincoln and Sheffield Motors Inc., two of the four Applewood Auto Group dealerships, were both open 24 days last month. Listed below is the number of vehicles serviced last month at the two dealerships. Construct dot plots and report summary statistics to compare the two dealerships.

| Tionesta Ford Lincoln |         |           |          |        |          |
|-----------------------|---------|-----------|----------|--------|----------|
| Monday                | Tuesday | Wednesday | Thursday | Friday | Saturday |
| 23                    | 33      | 27        | 28       | 39     | 26       |
| 30                    | 32      | 28        | 33       | 35     | 32       |
| 29                    | 25      | 36        | 31       | 32     | 27       |
| 35                    | 32      | 35        | 37       | 36     | 30       |

| Sheffield Motors Inc. |         |           |          |        |          |
|-----------------------|---------|-----------|----------|--------|----------|
| Monday                | Tuesday | Wednesday | Thursday | Friday | Saturday |
| 31                    | 35      | 44        | 36       | 34     | 37       |
| 30                    | 37      | 43        | 31       | 40     | 31       |
| 32                    | 44      | 36        | 34       | 43     | 36       |
| 26                    | 38      | 37        | 30       | 42     | 33       |

## SOLUTION

The Minitab system provides a dot plot and outputs the mean, median, maximum, and minimum values, and the standard deviation for the number of cars serviced at each dealership over the last 24 working days.



Source: Minitab

The dot plots, shown in the center of the output, graphically illustrate the distributions for each dealership. The plots show the difference in the location and dispersion of the observations. By looking at the dot plots, we can see that the number of vehicles serviced at the Sheffield dealership is more widely dispersed and has a larger mean than at the Tionesta dealership. Several other features of the number of vehicles serviced are:

- Tionesta serviced the fewest cars in any day, 23.
- Sheffield serviced 26 cars during their slowest day, which is 4 cars less than the next lowest day.
- Tionesta serviced exactly 32 cars on four different days.
- The numbers of cars serviced cluster around 36 for Sheffield and 32 for Tionesta.

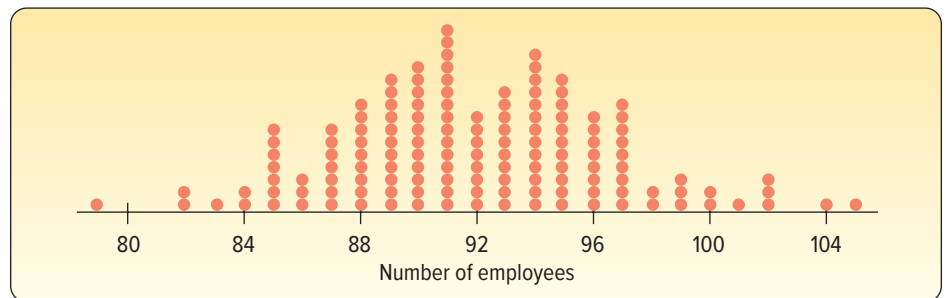
From the descriptive statistics, we see Sheffield serviced a mean of 35.83 vehicles per day. Tionesta serviced a mean of 31.292 vehicles per day during the same period. So Sheffield typically services 4.54 more vehicles per day. There is also more dispersion, or variation, in the daily number of vehicles serviced at Sheffield than at Tionesta. How do we know this? The standard deviation is larger at Sheffield (4.96 vehicles per day) than at Tionesta (4.112 cars per day).

**SELF-REVIEW 4-1**



©Steve Hix/Getty Images RF

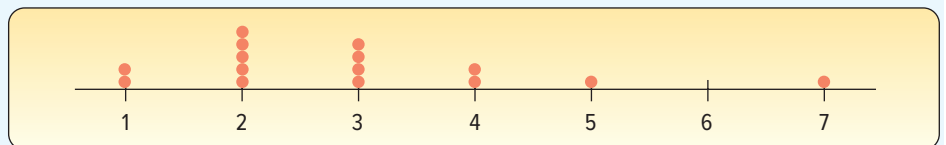
The number of employees at each of the 142 Home Depot stores in the Southeast region is shown in the following dot plot.



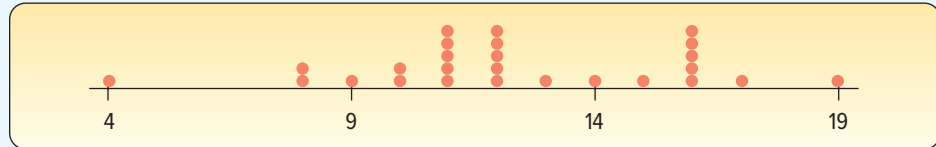
- (a) What are the maximum and minimum numbers of employees per store?
- (b) How many stores employ 91 people?
- (c) Around what values does the number of employees per store tend to cluster?

**EXERCISES**

- 1. Describe the differences between a histogram and a dot plot. When might a dot plot be better than a histogram?
- 2. When are dot plots most useful?
- 3. Consider the following chart.



- a. What is this chart called?
  - b. How many observations are in the study?
  - c. What are the maximum and the minimum values?
  - d. Around what values do the observations tend to cluster?
4. The following chart reports the number of cell phones sold at a big-box retail store for the last 26 days.



- a. What are the maximum and the minimum numbers of cell phones sold in a day?
- b. What is a typical number of cell phones sold?

**LO4-2**

Identify and compute measures of position.

## MEASURES OF POSITION

The standard deviation is the most widely used measure of dispersion. However, there are other ways of describing the variation or spread in a set of data. One method is to determine the *location* of values that divide a set of observations into equal parts. These measures include **quartiles**, **deciles**, and **percentiles**.

Quartiles divide a set of observations into four equal parts. To explain further, think of any set of values arranged from the minimum to the maximum. In Chapter 3, we called the middle value of a set of data arranged from the minimum to the maximum the median. That is, 50% of the observations are larger than the median and 50% are smaller. The median is a measure of location because it pinpoints the center of the data. In a similar fashion, **quartiles** divide a set of observations into four equal parts. The first quartile, usually labeled  $Q_1$ , is the value below which 25% of the observations occur, and the third quartile, usually labeled  $Q_3$ , is the value below which 75% of the observations occur.

Similarly, **deciles** divide a set of observations into 10 equal parts and **percentiles** into 100 equal parts. So if you found that your GPA was in the 8th decile at your university, you could conclude that 80% of the students had a GPA lower than yours and 20% had a higher GPA. If your GPA was in the 92nd percentile, then 92% of students had a GPA less than your GPA and only 8% of students had a GPA greater than your GPA. Percentile scores are frequently used to report results on such national standardized tests as the SAT, ACT, GMAT (used to judge entry into many master of business administration programs), and LSAT (used to judge entry into law school).

**QUARTILES** Values of an ordered (minimum to maximum) data set that divide the data into four intervals.

**DECILES** Values of an ordered (minimum to maximum) data set that divide the data into 10 equal parts.

**PERCENTILES** Values of an ordered (minimum to maximum) data set that divide the data into 100 intervals.

### Quartiles, Deciles, and Percentiles

To formalize the computational procedure, let  $L_p$  refer to the location of a desired percentile. So if we want to find the 92nd percentile we would use  $L_{92}$ , and if we wanted the median, the 50th percentile, then  $L_{50}$ . For a number of observations,  $n$ , the location of the  $P$ th percentile, can be found using the formula:

**LOCATION OF A PERCENTILE**

$$L_p = (n + 1) \frac{P}{100} \quad (4-1)$$

An example will help to explain further.

**EXAMPLE**

Morgan Stanley is an investment company with offices located throughout the United States. Listed below are the commissions earned last month by a sample of 15 brokers at the Morgan Stanley office in Oakland, California.

|         |         |         |         |         |         |         |         |         |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| \$2,038 | \$1,758 | \$1,721 | \$1,637 | \$2,097 | \$2,047 | \$2,205 | \$1,787 | \$2,287 |
| 1,940   | 2,311   | 2,054   | 2,406   | 1,471   | 1,460   |         |         |         |

Locate the median, the first quartile, and the third quartile for the commissions earned.

**SOLUTION**

The first step is to sort the data from the smallest commission to the largest.

|         |         |         |         |         |         |         |         |
|---------|---------|---------|---------|---------|---------|---------|---------|
| \$1,460 | \$1,471 | \$1,637 | \$1,721 | \$1,758 | \$1,787 | \$1,940 | \$2,038 |
| 2,047   | 2,054   | 2,097   | 2,205   | 2,287   | 2,311   | 2,406   |         |



©Yuji Kotani/Getty Images RF

The median value is the observation in the center and is the same as the 50th percentile, so  $P$  equals 50. So the median or  $L_{50}$  is located at  $(n + 1)(50/100)$ , where  $n$  is the number of observations. In this case, that is position number 8, found by  $(15 + 1)(50/100)$ . The eighth-largest commission is \$2,038. So we conclude this is the median and that half the brokers earned commissions more than \$2,038 and half earned less than \$2,038. The result using formula

(4-1) to find the median is the same as the method presented in Chapter 3.

Recall the definition of a quartile. Quartiles divide a set of observations into four equal parts. Hence 25% of the observations will be less than the first quartile. Seventy-five percent of the observations will be less than the third quartile. To locate the first quartile, we use formula (4-1), where  $n = 15$  and  $P = 25$ :

$$L_{25} = (n + 1) \frac{P}{100} = (15 + 1) \frac{25}{100} = 4$$

And to locate the third quartile,  $n = 15$  and  $P = 75$ :

$$L_{75} = (n + 1) \frac{P}{100} = (15 + 1) \frac{75}{100} = 12$$

Therefore, the first and third quartile values are located at positions 4 and 12, respectively. The fourth value in the ordered array is \$1,721 and the twelfth is \$2,205. These are the first and third quartiles.

In the above example, the location formula yielded a whole number. That is, we wanted to find the first quartile and there were 15 observations, so the location formula indicated we should find the fourth ordered value. What if there were 20 observations in the sample, that is  $n = 20$ , and we wanted to locate the first quartile? From the location formula (4–1):

$$L_{25} = (n + 1)\frac{P}{100} = (20 + 1)\frac{25}{100} = 5.25$$

We would locate the fifth value in the ordered array and then move .25 of the distance between the fifth and sixth values and report that as the first quartile. Like the median, the quartile does not need to be one of the actual values in the data set.

To explain further, suppose a data set contained the six values 91, 75, 61, 101, 43, and 104. We want to locate the first quartile. We order the values from the minimum to the maximum: 43, 61, 75, 91, 101, and 104. The first quartile is located at

$$L_{25} = (n + 1)\frac{P}{100} = (6 + 1)\frac{25}{100} = 1.75$$

The position formula tells us that the first quartile is located between the first and the second values and it is .75 of the distance between the first and the second values. The first value is 43 and the second is 61. So the distance between these two values is 18. To locate the first quartile, we need to move .75 of the distance between the first and second values, so  $.75(18) = 13.5$ . To complete the procedure, we add 13.5 to the first value, 43, and report that the first quartile is 56.5.

We can extend the idea to include both deciles and percentiles. To locate the 23rd percentile in a sample of 80 observations, we would look for the 18.63 position.

$$L_{23} = (n + 1)\frac{P}{100} = (80 + 1)\frac{23}{100} = 18.63$$

To find the value corresponding to the 23rd percentile, we would locate the 18th value and the 19th value and determine the distance between the two values. Next, we would multiply this difference by 0.63 and add the result to the smaller value. The result would be the 23rd percentile.

Statistical software is very helpful when describing and summarizing data. Excel, Minitab, and MegaStat, a statistical analysis Excel add-in, all provide summary statistics that include quartiles. For example, the Minitab summary of the Morgan Stanley commission data, shown below, includes the first and third quartiles and other statistics. Based on the reported quartiles, 25% of the commissions earned were less than \$1,721 and 75% were less than \$2,205. These are the same values we calculated using formula (4–1).

#### STATISTICS IN ACTION

John W. Tukey (1915–2000) received a PhD in mathematics from Princeton University in 1939. However, when he joined the Fire Control Research Office during World War II, his interest in abstract mathematics shifted to applied statistics. He developed effective numerical and graphical methods for studying patterns in data. Among the graphics he developed is the box-and-whisker plot or box plot. From 1960 to 1980, Tukey headed the statistical division of NBC's election night vote projection team. He became renowned in 1960 for preventing an early call of victory for Richard Nixon in the presidential election won by John F. Kennedy.

The screenshot shows a Minitab window with a data table in column C1 and a 'Descriptive Statistics: Commissions' window. The data table has 4 rows of commission values. The Descriptive Statistics window displays the following summary statistics:

| Variable    | N  | N* | Mean   | SE Mean | StDev | Minimum | Q1     | Median | Q3     | Maximum |
|-------------|----|----|--------|---------|-------|---------|--------|--------|--------|---------|
| Commissions | 15 | 0  | 1947.9 | 77.1    | 298.8 | 1460.0  | 1721.0 | 2038.0 | 2205.0 | 2406.0  |

Source: Minitab

| Morgan Stanley Commissions |                  |        |
|----------------------------|------------------|--------|
| 1460                       | Equation 4-1     |        |
| 2047                       | Quartile 1       | 1721   |
| 1471                       | Quartile 3       | 2205   |
| 2054                       |                  |        |
| 1637                       |                  |        |
| 2097                       | Alternate Method |        |
| 1721                       | Quartile 1       | 1739.5 |
| 2205                       | Quartile 3       | 2151   |
| 1758                       |                  |        |
| 2287                       |                  |        |
| 1787                       |                  |        |
| 2311                       |                  |        |
| 1940                       |                  |        |
| 2406                       |                  |        |
| 2038                       |                  |        |

There are ways other than formula (4–1) to locate quartile values. For example, another method uses  $0.25n + 0.75$  to locate the position of the first quartile and  $0.75n + 0.25$  to locate the position of the third quartile. We will call this the *Excel Method*. In the Morgan Stanley data, this method would place the first quartile at position 4.5 ( $.25 \times 15 + .75$ ) and the third quartile at position 11.5 ( $.75 \times 15 + .25$ ). The first quartile would be interpolated as 0.5, or one-half the difference between the fourth- and the fifth-ranked values. Based on this method, the first quartile is \$1739.5, found by  $(\$1,721 + 0.5[\$1,758 - \$1,721])$ . The third quartile, at position 11.5, would be \$2,151, or one-half the distance between the eleventh- and the

twelfth-ranked values, found by  $(\$2,097 + 0.5[\$2,205 - \$2,097])$ . Excel, as shown in the Morgan Stanley and Applewood examples, can compute quartiles using either of the two methods. **Please note the text uses formula (4–1) to calculate quartiles.**

Is the difference between the two methods important? No. Usually it is just a nuisance. In general, both methods calculate values that will support the statement that approximately 25% of the values are less than the value of the first quartile, and approximately 75% of the data values are less than the value of the third quartile. When the sample is large, the difference in the results from the two methods is small. For example, in the Applewood Auto Group data there are 180 vehicles. The quartiles computed using both methods are shown to the left. Based on the variable profit, 45 of the 180 values (25%) are less than both values of the first quartile, and 135 of the 180 values (75%) are less than both values of the third quartile.

| Applewood |         |                  |        |
|-----------|---------|------------------|--------|
| Age       | Profit  |                  |        |
| 21        | \$1,387 |                  |        |
| 23        | \$1,754 |                  |        |
| 24        | \$1,817 | Equation 4-1     |        |
| 25        | \$1,040 | Quartile 1       | 1415.5 |
| 26        | \$1,273 | Quartile 3       | 2275.5 |
| 27        | \$1,529 |                  |        |
| 27        | \$3,082 | Alternate Method |        |
| 28        | \$1,951 | Quartile 1       | 1422.5 |
| 28        | \$2,692 | Quartile 3       | 2268.5 |
| 29        | \$1,342 |                  |        |

When using Excel, be careful to understand the method used to calculate quartiles. Excel 2013 and Excel 2016 offer both methods. The Excel function, **Quartile.exc**, will result in the same answer as Equation 4–1. The Excel function, **Quartile.inc**, will result in the Excel Method answers.

## SELF-REVIEW 4-2



The Quality Control department of Plainsville Peanut Company is responsible for checking the weight of the 8-ounce jar of peanut butter. The weights of a sample of nine jars produced last hour are:

|      |      |      |      |      |      |      |      |      |
|------|------|------|------|------|------|------|------|------|
| 7.69 | 7.72 | 7.80 | 7.86 | 7.90 | 7.94 | 7.97 | 8.06 | 8.09 |
|------|------|------|------|------|------|------|------|------|

- What is the median weight?
- Determine the weights corresponding to the first and third quartiles.



## EXERCISES

5. **FILE** Determine the median and the first and third quartiles in the following data.

46 47 49 49 51 53 54 54 55 55 59

6. **FILE** Determine the median and the first and third quartiles in the following data.

5.24 6.02 6.67 7.30 7.59 7.99 8.03 8.35 8.81 9.45  
9.61 10.37 10.39 11.86 12.22 12.71 13.07 13.59 13.89 15.42

7. **FILE** The Thomas Supply Company Inc. is a distributor of gas-powered generators. As with any business, the length of time customers take to pay their invoices is important. Listed below, arranged from smallest to largest, is the time, in days, for a sample of the Thomas Supply Company Inc. invoices.

13 13 13 20 26 27 31 34 34 34 35 35 36 37 38  
41 41 41 45 47 47 47 50 51 53 54 56 62 67 82

- a. Determine the first and third quartiles.  
b. Determine the second decile and the eighth decile.  
c. Determine the 67th percentile.
8. **FILE** Kevin Horn is the national sales manager for National Textbooks Inc. He has a sales staff of 40 who visit college professors all over the United States. Each Saturday morning, he requires his sales staff to send him a report. This report includes, among other things, the number of professors visited during the previous week. Listed below, ordered from smallest to largest, are the number of visits last week.

38 40 41 45 48 48 50 50 51 51 52 52 53 54 55 55 55 56 56 57  
59 59 59 62 62 62 63 64 65 66 66 67 67 69 69 71 77 78 79 79

- a. Determine the median number of calls.  
b. Determine the first and third quartiles.  
c. Determine the first decile and the ninth decile.  
d. Determine the 33rd percentile.

**LO4-3**

Construct and analyze a box plot.

## BOX PLOTS

A **box plot** is a graphical display, based on quartiles, that helps us picture a set of data. To construct a box plot, we need only five statistics: the minimum value,  $Q_1$  (the first quartile), the median,  $Q_3$  (the third quartile), and the maximum value. An example will help to explain.

**BOX PLOT** A graphic display that shows the general shape of a variable's distribution. It is based on five descriptive statistics: the maximum and minimum values, the first and third quartiles, and the median.

**EXAMPLE**

Alexander's Pizza offers free delivery of its pizza within 15 miles. Alex, the owner, wants some information on the time it takes for delivery. How long does a typical delivery take? Within what range of times will most deliveries be completed? For a sample of 20 deliveries, he determined the following information:

Minimum value = 13 minutes

$Q_1$  = 15 minutes

Median = 18 minutes

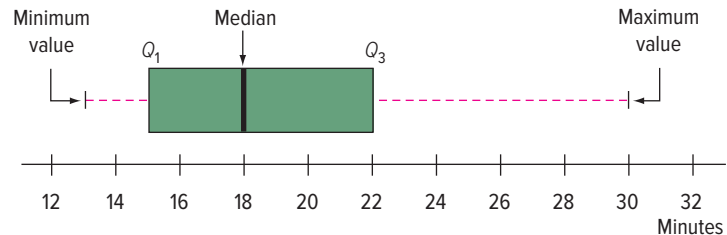
$Q_3$  = 22 minutes

Maximum value = 30 minutes

Develop a box plot for the delivery times. What conclusions can you make about the delivery times?

**SOLUTION**

The first step in drawing a box plot is to create an appropriate scale along the horizontal axis. Next, we draw a box that starts at  $Q_1$  (15 minutes) and ends at  $Q_3$  (22 minutes). Inside the box we place a vertical line to represent the median (18 minutes). Finally, we extend horizontal lines from the box out to the minimum value (13 minutes) and the maximum value (30 minutes). These horizontal lines outside of the box are sometimes called “whiskers” because they look a bit like a cat’s whiskers.



The box plot also shows the interquartile range of delivery times between  $Q_1$  and  $Q_3$ . The **interquartile range** is 7 minutes and indicates that 50% of the deliveries are between 15 and 22 minutes.

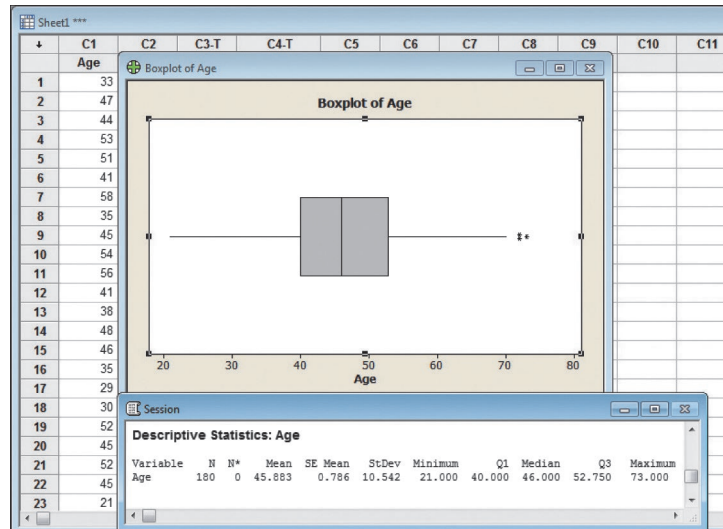
The box plot also reveals that the distribution of delivery times is positively skewed. In Chapter 3, we defined skewness as the lack of symmetry in a set of data. How do we know this distribution is positively skewed? In this case, there are actually two pieces of information that suggest this. First, the dashed line to the right of the box from 22 minutes ( $Q_3$ ) to the maximum time of 30 minutes is longer than the dashed line from the left of 15 minutes ( $Q_1$ ) to the minimum value of 13 minutes. To put it another way, the 25% of the data larger than the third quartile are more spread out than the 25% less than the first quartile. A second indication of positive skewness is that the median is not in the center of the box. The distance from the first quartile to the median is smaller than the distance from the median to the third quartile. We know that the number of delivery times between 15 minutes and 18 minutes is the same as the number of delivery times between 18 minutes and 22 minutes.

### EXAMPLE

Refer to the Applewood Auto Group data. Develop a box plot for the variable age of the buyer. What can we conclude about the distribution of the age of the buyer?

### SOLUTION

Minitab was used to develop the following chart and summary statistics.



Source: Minitab

The median age of the purchaser is 46 years, 25% of the purchasers are less than 40 years of age, and 25% are more than 52.75 years of age. Based on the summary information and the box plot, we conclude:

- Fifty percent of the purchasers are between the ages of 40 and 52.75 years.
- The distribution of ages is fairly symmetric. There are two reasons for this conclusion. The length of the whisker above 52.75 years ( $Q_3$ ) is about the same length as the whisker below 40 years ( $Q_1$ ). Also, the area in the box between 40 years and the median of 46 years is about the same as the area between the median and 52.75.

There are three asterisks (\*) above 70 years. What do they indicate? In a box plot, an asterisk identifies an **outlier**. An outlier is a value that is inconsistent with the rest of the data. It is defined as a value that is more than 1.5 times the interquartile range smaller than  $Q_1$  or larger than  $Q_3$ . In this example, an outlier would be a value larger than 71.875 years, found by:

$$\text{Outlier} > Q_3 + 1.5(Q_3 - Q_1) = 52.75 + 1.5(52.75 - 40) = 71.875$$

An outlier would also be a value less than 20.875 years.

$$\text{Outlier} < Q_1 - 1.5(Q_3 - Q_1) = 40 - 1.5(52.75 - 40) = 20.875$$

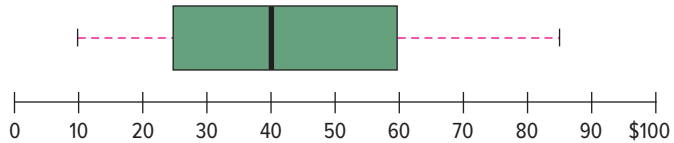
**OUTLIER** A data point that is unusually far from the others. An accepted rule is to classify an observation as an outlier if it is 1.5 times the interquartile range above the third quartile or below the first quartile.

From the box plot, we conclude there are three purchasers 72 years of age or older and none less than 21 years of age. Technical note: In some cases, a single asterisk may represent more than one observation because of the limitations of the software and space available. It is a good idea to check the actual data. In this instance, there are three purchasers 72 years old or older; two are 72 and one is 73.

### SELF-REVIEW 4-3



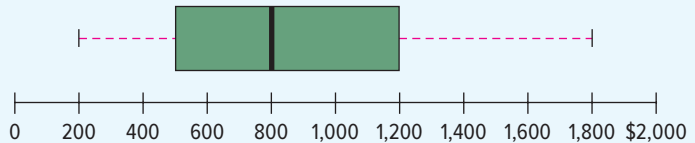
The following box plot shows the assets in millions of dollars for credit unions in Seattle, Washington.



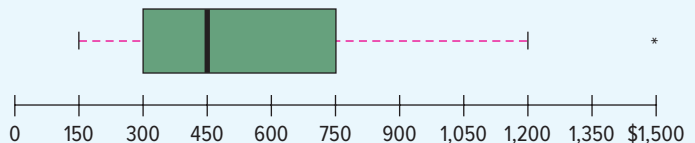
What are the smallest and largest values, the first and third quartiles, and the median? Would you agree that the distribution is symmetrical? Are there any outliers?

### EXERCISES

9. The box plot below shows the amount spent for books and supplies per year by students at four-year public colleges.



- Estimate the median amount spent.
  - Estimate the first and third quartiles for the amount spent.
  - Estimate the interquartile range for the amount spent.
  - Beyond what point is a value considered an outlier?
  - Identify any outliers and estimate their values.
  - Is the distribution symmetrical or positively or negatively skewed?
10. The box plot shows the undergraduate in-state tuition per credit hour at four-year public colleges.



- Estimate the median.
  - Estimate the first and third quartiles.
  - Determine the interquartile range.
  - Beyond what point is a value considered an outlier?
  - Identify any outliers and estimate their values.
  - Is the distribution symmetrical or positively or negatively skewed?
11. In a study of the gasoline mileage of model year 2017 automobiles, the mean miles per gallon was 27.5 and the median was 26.8. The smallest value in the study was 12.70 miles per gallon, and the largest was 50.20. The first and third quartiles were 17.95 and 35.45 miles per gallon, respectively. Develop a box plot and comment on the distribution. Is it a symmetric distribution?

12. **FILE** A sample of 28 time-shares in the Orlando, Florida, area revealed the following daily charges for a one-bedroom suite. For convenience, the data are ordered from smallest to largest. Construct a box plot to represent the data. Comment on the distribution. Be sure to identify the first and third quartiles and the median.

|       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|
| \$116 | \$121 | \$157 | \$192 | \$207 | \$209 | \$209 |
| 229   | 232   | 236   | 236   | 239   | 243   | 246   |
| 260   | 264   | 276   | 281   | 283   | 289   | 296   |
| 307   | 309   | 312   | 317   | 324   | 341   | 353   |

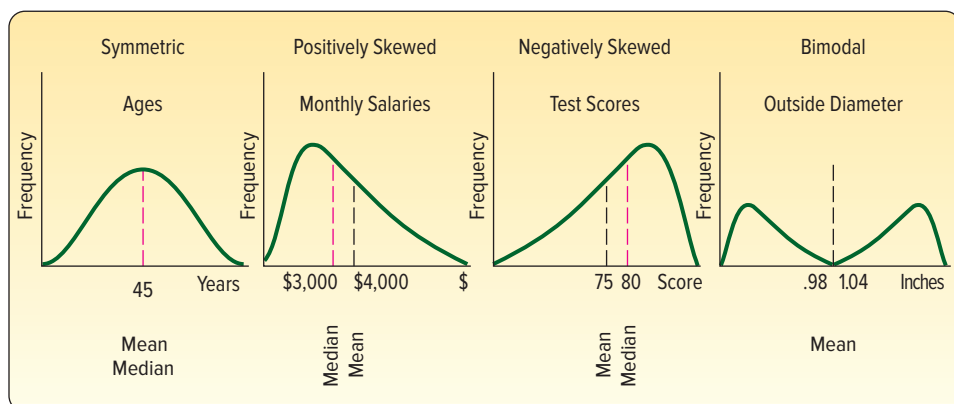
**LO4-4**

Compute and interpret the coefficient of skewness.

**SKEWNESS**

In Chapter 3, we described measures of central location for a distribution of data by reporting the mean, median, and mode. We also described measures that show the amount of spread or variation in a distribution, such as the range and the standard deviation.

Another characteristic of a distribution is the shape. There are four shapes commonly observed: symmetric, positively skewed, negatively skewed, and bimodal. In a **symmetric** distribution the mean and median are equal and the data values are evenly spread around these values. The shape of the distribution below the mean and median is a mirror image of distribution above the mean and median. A distribution of values is **skewed to the right** or **positively skewed** if there is a single peak, but the values extend much farther to the right of the peak than to the left of the peak. In this case, the mean is larger than the median. In a **negatively skewed** distribution there is a single peak, but the observations extend farther to the left, in the negative direction, than to the right. In a negatively skewed distribution, the mean is smaller than the median. Positively skewed distributions are more common. Salaries often follow this pattern. Think of the salaries of those employed in a small company of about 100 people. The president and a few top executives would have very large salaries relative to the other workers and hence the distribution of salaries would exhibit positive skewness. A **bimodal distribution** will have two or more peaks. This is often the case when the values are from two or more populations. This information is summarized in Chart 4–1.



**CHART 4–1** Shapes of Frequency Polygons

There are several formulas in the statistical literature used to calculate skewness. The simplest, developed by Professor Karl Pearson (1857–1936), is based on the difference between the mean and the median.

**STATISTICS IN ACTION**

The late Stephen Jay Gould (1941–2002) was a professor of zoology and professor of geology at Harvard University. In 1982, he was diagnosed with cancer and had an expected survival time of 8 months. However, never one to be discouraged, his research showed that the distribution of survival time is dramatically skewed to the right and showed that not only do 50% of similar cancer patients survive more than 8 months, but that the survival time could be years rather than months! In fact, Dr. Gould lived another 20 years. Based on his experience, he wrote a widely published essay titled “The Median Isn’t the Message.”

**PEARSON’S COEFFICIENT OF SKEWNESS**

$$sk = \frac{3(\bar{x} - \text{Median})}{s} \quad (4-2)$$

Using this relationship, the coefficient of skewness can range from  $-3$  up to  $3$ . A value near  $-3$ , such as  $-2.57$ , indicates considerable negative skewness. A value such as  $1.63$  indicates moderate positive skewness. A value of  $0$ , which will occur when the mean and median are equal, indicates the distribution is symmetrical and there is no skewness present.

In this text, we present output from Minitab and Excel. Both of these software packages compute a value for the coefficient of skewness based on the cubed deviations from the mean. The formula is:

**SOFTWARE COEFFICIENT OF SKEWNESS**

$$sk = \frac{n}{(n-1)(n-2)} \left[ \sum \left( \frac{x - \bar{x}}{s} \right)^3 \right] \quad (4-3)$$

Formula (4–3) offers an insight into skewness. The right-hand side of the formula is the difference between each value and the mean, divided by the standard deviation. That is the portion  $(x - \bar{x})/s$  of the formula. This idea is called **standardizing**. We will discuss the idea of standardizing a value in more detail in Chapter 7 when we describe the normal probability distribution. At this point, observe that the result is to report the difference between each value and the mean in units of the standard deviation. If this difference is positive, the particular value is larger than the mean; if the value is negative, the standardized quantity is smaller than the mean. When we cube these values, we retain the information on the direction of the difference. Recall that in the formula for the standard deviation [see formula (3–8)], we squared the difference between each value and the mean, so that the result was all nonnegative values.

If the set of data values under consideration is symmetric, when we cube the standardized values and sum over all the values, the result will be near zero. If there are several large values, clearly separate from the others, the sum of the cubed differences will be a large positive value. If there are several small values clearly separate from the others, the sum of the cubed differences will be negative.

An example will illustrate the idea of skewness.

**EXAMPLE**

Following are the earnings per share for a sample of 15 software companies for the year 2017. The earnings per share are arranged from smallest to largest.

|        |        |        |        |         |         |         |        |
|--------|--------|--------|--------|---------|---------|---------|--------|
| \$0.09 | \$0.13 | \$0.41 | \$0.51 | \$ 1.12 | \$ 1.20 | \$ 1.49 | \$3.18 |
| 3.50   | 6.36   | 7.83   | 8.92   | 10.13   | 12.99   | 16.40   |        |

Compute the mean, median, and standard deviation. Find the coefficient of skewness using Pearson’s estimate and the software methods. What is your conclusion regarding the shape of the distribution?

**SOLUTION**

These are sample data, so we use formula (3–2) to determine the mean.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{\$74.26}{15} = \$4.95$$

The median is the middle value in a set of data, arranged from smallest to largest. In this case, there is an odd number of observations, so the middle value is the median. It is \$3.18.

We use formula (3–8) on page 77 to determine the sample standard deviation.

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{(\$0.09 - \$4.95)^2 + \dots + (\$16.40 - \$4.95)^2}{15 - 1}} = \$5.22$$

Pearson's coefficient of skewness is 1.017, found by

$$sk = \frac{3(\bar{x} - \text{Median})}{s} = \frac{3(\$4.95 - \$3.18)}{\$5.22} = 1.017$$

This indicates there is moderate positive skewness in the earnings per share data.

We obtain a similar, but not exactly the same, value from the software method. The details of the calculations are shown in Table 4–1. To begin, we find the difference between each earnings per share value and the mean and divide this result by the standard deviation. We have referred to this as standardizing. Next, we cube, that is, raise to the third power, the result of the first step. Finally, we sum the cubed values. The details for the first company, that is, the company with an earnings per share of \$0.09, are:

$$\left(\frac{x - \bar{x}}{s}\right)^3 = \left(\frac{0.09 - 4.95}{5.22}\right)^3 = (-0.9310)^3 = -0.8070$$

**TABLE 4–1** Calculation of the Coefficient of Skewness

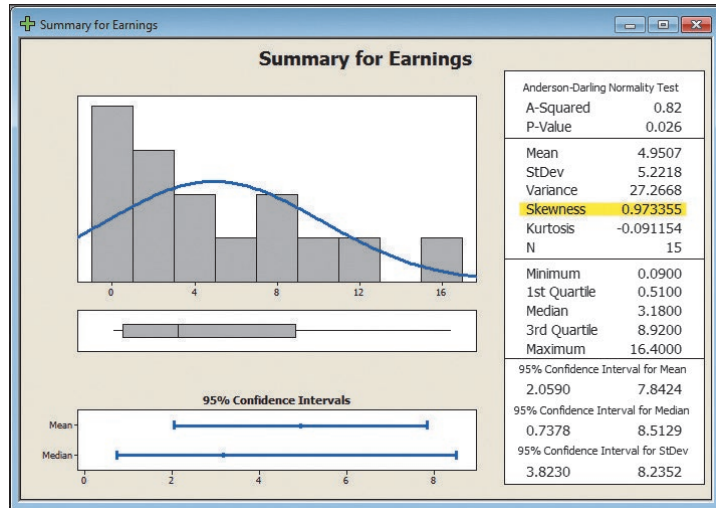
| Earnings per Share | $\frac{(x - \bar{x})}{s}$ | $\left(\frac{x - \bar{x}}{s}\right)^3$ |
|--------------------|---------------------------|--|
| 0.09               | −0.9310                   | −0.8070                                |
| 0.13               | −0.9234                   | −0.7873                                |
| 0.41               | −0.8697                   | −0.6579                                |
| 0.51               | −0.8506                   | −0.6154                                |
| 1.12               | −0.7337                   | −0.3950                                |
| 1.20               | −0.7184                   | −0.3708                                |
| 1.49               | −0.6628                   | −0.2912                                |
| 3.18               | −0.3391                   | −0.0390                                |
| 3.50               | −0.2778                   | −0.0214                                |
| 6.36               | 0.2701                    | 0.0197                                 |
| 7.83               | 0.5517                    | 0.1679                                 |
| 8.92               | 0.7605                    | 0.4399                                 |
| 10.13              | 0.9923                    | 0.9772                                 |
| 12.99              | 1.5402                    | 3.6539                                 |
| 16.40              | 2.1935                    | 10.5537                                |
|                    |                           | <u>11.8274</u>                         |

When we sum the 15 cubed values, the result is 11.8274. That is, the term  $\sum[(x - \bar{x})/s]^3 = 11.8274$ . To find the coefficient of skewness, we use formula (4–3), with  $n = 15$ .

$$sk = \frac{n}{(n - 1)(n - 2)} \sum \left(\frac{x - \bar{x}}{s}\right)^3 = \frac{15}{(15 - 1)(15 - 2)} (11.8274) = 0.975$$

We conclude that the earnings per share values are somewhat positively skewed. The following Minitab summary reports the descriptive measures, such as

the mean, median, and standard deviation of the earnings per share data. Also included are the coefficient of skewness and a histogram with a bell-shaped curve superimposed.



Source: Minitab

### SELF-REVIEW 4-4



A sample of five data entry clerks employed in the Horry County Tax Office revised the following number of tax records last hour: 73, 98, 60, 92, and 84.

- Find the mean, median, and the standard deviation.
- Compute the coefficient of skewness using Pearson's method.
- Calculate the coefficient of skewness using the software method.
- What is your conclusion regarding the skewness of the data?

### EXERCISES

For Exercises 13–16:

- Determine the mean, median, and the standard deviation.
- Determine the coefficient of skewness using Pearson's method.
- Determine the coefficient of skewness using the software method.

13. **FILE** The following values are the starting salaries, in \$000, for a sample of five accounting graduates who accepted positions in public accounting last year.

36.0    26.0    33.0    28.0    31.0

14. **FILE** Listed below are the salaries, in \$000, for a sample of 15 chief financial officers in the electronics industry.

\$516.0    \$548.0    \$566.0    \$534.0    \$586.0    \$529.0  
 546.0    523.0    538.0    523.0    551.0    552.0  
 486.0    558.0    574.0



15. **FILE** Listed below are the commissions earned (\$000) last year by the 15 sales representatives at Furniture Patch Inc.

|        |        |        |        |        |        |        |        |        |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| \$ 3.9 | \$ 5.7 | \$ 7.3 | \$10.6 | \$13.0 | \$13.6 | \$15.1 | \$15.8 | \$17.1 |
| 17.4   | 17.6   | 22.3   | 38.6   | 43.2   | 87.7   |        |        |        |

16. **FILE** Listed below are the salaries for the 2016 New York Yankees Major League Baseball team.

| Player          | Salary       | Player          | Salary      |
|-----------------|--------------|-----------------|-------------|
| CC Sabathia     | \$25,000,000 | Dustin Ackley   | \$3,200,000 |
| Mark Teixeira   | 23,125,000   | Martin Prado    | 3,000,000   |
| Masahiro Tanaka | 22,000,000   | Didi Gregorius  | 2,425,000   |
| Jacoby Ellsbury | 21,142,857   | Aaron Hicks     | 574,000     |
| Alex Rodriguez  | 21,000,000   | Austin Romine   | 556,000     |
| Brian McCann    | 17,000,000   | Chasen Shreve   | 533,400     |
| Carlos Beltran  | 15,000,000   | Greg Bird       | 525,300     |
| Brett Gardner   | 13,500,000   | Luis Severino   | 521,300     |
| Chase Headley   | 13,000,000   | Bryan Mitchell  | 516,650     |
| Aroldis Chapman | 11,325,000   | Kirby Yates     | 511,900     |
| Andrew Miller   | 9,000,000    | Mason Williams  | 509,700     |
| Starlin Castro  | 7,857,143    | Ronald Torreyes | 508,600     |
| Nathan Eovaldi  | 5,600,000    | John Barbatto   | 507,500     |
| Michael Pineda  | 4,300,000    | Dellin Betances | 507,500     |
| Ivan Nova       | 4,100,000    | Luis Cessa      | 507,500     |

#### LO4-5

Create and interpret a scatter diagram.

## DESCRIBING THE RELATIONSHIP BETWEEN TWO VARIABLES

In Chapter 2 and the first section of this chapter, we presented graphical techniques to summarize the distribution of a single variable. We used a histogram in Chapter 2 to summarize the profit on vehicles sold by the Applewood Auto Group. Earlier in this chapter, we used dot plots to visually summarize a set of data. Because we are studying a single variable, we refer to this as **univariate** data.



©Steve Mason/Getty Images RF

There are situations where we wish to study and visually portray the relationship between two variables. When we study the relationship between two variables, we refer to the data as **bivariate**. Data analysts frequently wish to understand the relationship between two variables. Here are some examples:

- Tybo and Associates is a law firm that advertises extensively on local TV. The partners are considering increasing their advertising budget. Before doing so, they would like to know the relationship between the amount spent per month on advertising and the total amount of billings for that month. To put it another way, will increasing the amount spent on advertising result in an increase in billings?

- Coastal Realty is studying the selling prices of homes. What variables seem to be related to the selling price of homes? For example, do larger homes sell for more than smaller ones? Probably. So Coastal might study the relationship between the area in square feet and the selling price.
- Dr. Stephen Givens is an expert in human development. He is studying the relationship between the height of fathers and the height of their sons. That is, do tall fathers tend to have tall children? Would you expect LeBron James, the 6'8", 250-pound professional basketball player, to have relatively tall sons?

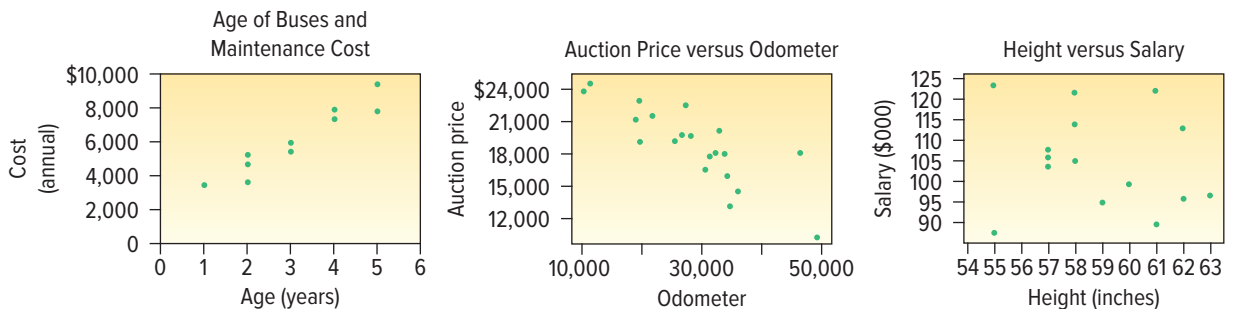
One graphical technique we use to show the relationship between variables is called a **scatter diagram**.

**SCATTER DIAGRAM** Graphical technique used to show the relationship between two variables measured with interval or ratio scales.

To draw a scatter diagram, we need two variables. We scale one variable along the horizontal axis ( $X$ -axis) of a graph and the other variable along the vertical axis ( $Y$ -axis). Usually one variable depends to some degree on the other. In the third example above, the height of the son *depends* on the height of the father. So we scale the height of the father on the horizontal axis and that of the son on the vertical axis.

We can use statistical software, such as Excel, to perform the plotting function for us. *Caution:* You should always be careful of the scale. By changing the scale of either the vertical or the horizontal axis, you can affect the apparent visual strength of the relationship.

Following are three scatter diagrams (Chart 4–2). The one on the left shows a rather strong positive relationship between the age in years and the maintenance cost last year for a sample of 10 buses owned by the city of Cleveland, Ohio. Note that as the age of the bus increases, the yearly maintenance cost also increases. The example in the center, for a sample of 20 vehicles, shows a rather strong indirect relationship between the odometer reading and the auction price. That is, as the number of miles driven increases, the auction price decreases. The example on the right depicts the relationship between the height and yearly salary for a sample of 15 shift supervisors. This graph indicates there is little relationship between their height and yearly salary.



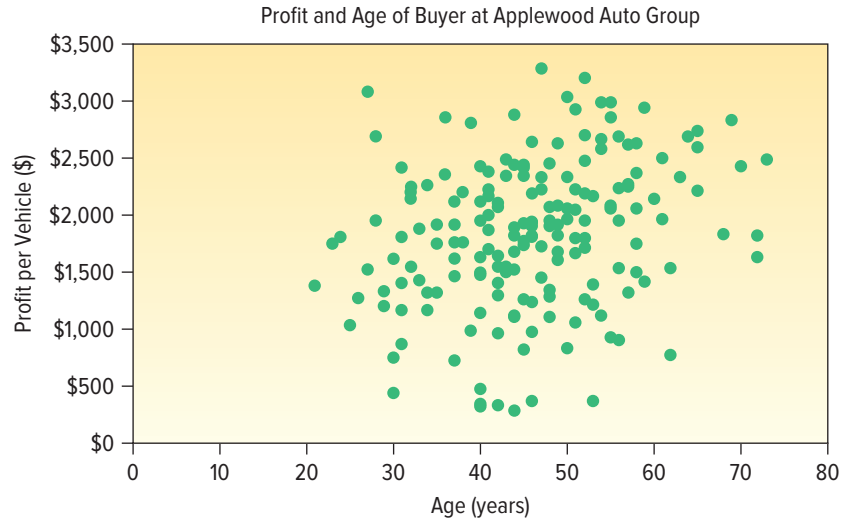
**CHART 4–2** Three Examples of Scatter Diagrams

### ▶ EXAMPLE

In the introduction to Chapter 2, we presented data from the Applewood Auto Group. We gathered information concerning several variables, including the profit earned from the sale of 180 vehicles sold last month. In addition to the amount of profit on each sale, one of the other variables is the age of the purchaser. Is there a relationship between the profit earned on a vehicle sale and the age of the purchaser? Would it be reasonable to conclude that more profit is made on vehicles purchased by older buyers?

## SOLUTION

We can investigate the relationship between vehicle profit and the age of the buyer with a scatter diagram. We scale age on the horizontal, or  $X$ -axis, and the profit on the vertical, or  $Y$ -axis. We assume profit depends on the age of the purchaser. As people age, they earn more income and purchase more expensive cars which, in turn, produce higher profits. We use Excel to develop the scatter diagram. The Excel commands are in Appendix C.



The scatter diagram shows a rather weak positive relationship between the two variables. It does not appear there is much relationship between the vehicle profit and the age of the buyer. In Chapter 13, we will study the relationship between variables more extensively, even calculating several numerical measures to express the relationship between variables.

In the preceding example, there is a weak positive, or direct, relationship between the variables. There are, however, many instances where there is a relationship between the variables, but that relationship is inverse or negative. For example:

- The value of a vehicle and the number of miles driven. As the number of miles increases, the value of the vehicle decreases.
- The premium for auto insurance and the age of the driver. Auto rates tend to be the highest for younger drivers and lower for older drivers.
- For many law enforcement personnel, as the number of years on the job increases, the number of traffic citations decreases. This may be because personnel become more liberal in their interpretations or they may be in supervisor positions and not in a position to issue as many citations. But in any event, as age increases, the number of citations decreases.

### LO4-6

Develop and explain a contingency table.

## CONTINGENCY TABLES

A scatter diagram requires that both of the variables be at least interval scale. In the Applewood Auto Group example, both age and vehicle profit are ratio-scale variables. Height is also ratio scale as used in the discussion of the relationship between the height of fathers and the height of their sons. What if we wish to study the relationship between two variables when one or both are nominal or ordinal scale? In this case, we tally the results in a **contingency table**.

**CONTINGENCY TABLE** A table used to classify sample observations according to two identifiable characteristics.

A contingency table is a cross-tabulation that simultaneously summarizes two variables of interest. For example:

- Students at a university are classified by gender and class (freshman, sophomore, junior, or senior).
- A product is classified as acceptable or unacceptable and by the shift (day, afternoon, or night) on which it is manufactured.
- A voter in a school bond referendum is classified as to party affiliation (Democrat, Republican, other) and the number of children that voter has attending school in the district (0, 1, 2, etc.).

### EXAMPLE

There are four dealerships in the Applewood Auto Group. Suppose we want to compare the profit earned on each vehicle sold by the particular dealership. To put it another way, is there a relationship between the amount of profit earned and the dealership?

### SOLUTION

In a contingency table, both variables only need to be nominal or ordinal. In this example, the variable dealership is a nominal variable and the variable profit is a ratio variable. To convert profit to an ordinal variable, we classify the variable profit into two categories, those cases where the profit earned is more than the median and those cases where it is less. On page 66, we calculated the median profit for all sales last month at Applewood Auto Group to be \$1,882.50.

| Above/Below Median Profit | Kane | Olean | Sheffield | Tionesta | Total |
|---------------------------|------|-------|-----------|----------|-------|
| Above                     | 25   | 20    | 19        | 26       | 90    |
| Below                     | 27   | 20    | 26        | 17       | 90    |
| Total                     | 52   | 40    | 45        | 43       | 180   |

By organizing the information into a contingency table, we can compare the profit at the four dealerships. We observe the following:

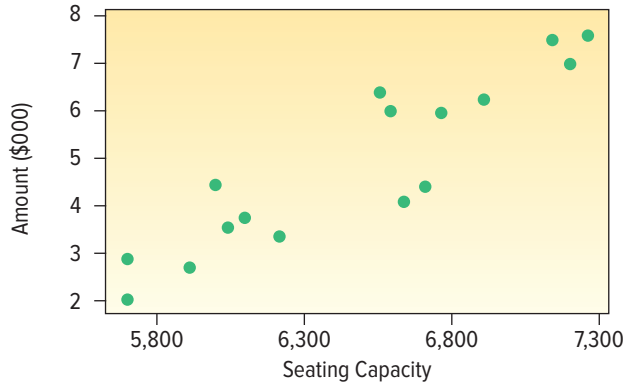
- From the Total column on the right, 90 of the 180 cars sold had a profit above the median and half below. From the definition of the median, this is expected.
- For the Kane dealership, 25 out of the 52, or 48%, of the cars sold were sold for a profit more than the median.
- The percentage of profits above the median for the other dealerships are 50% for Olean, 42% for Sheffield, and 60% for Tionesta.

We will return to the study of contingency tables in Chapter 5 during the study of probability and in Chapter 15 during the study of nonparametric methods of analysis.

**SELF-REVIEW 4-5**



The rock group Blue String Beans is touring the United States. The following chart shows the relationship between concert seating capacity and revenue in \$000 for a sample of concerts.



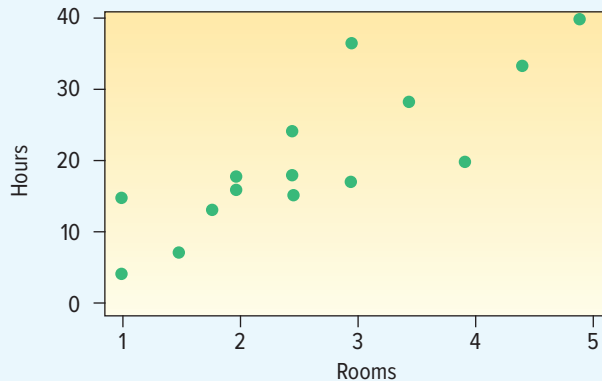
- (a) What is the diagram called?
- (b) How many concerts were studied?
- (c) Estimate the revenue for the concert with the largest seating capacity.
- (d) How would you characterize the relationship between revenue and seating capacity? Is it strong or weak, direct or inverse?

**EXERCISES**

17. **FILE** Develop a scatter diagram for the following sample data. How would you describe the relationship between the values?

| x-Value | y-Value | x-Value | y-Value |
|---------|---------|---------|---------|
| 10      | 6       | 11      | 6       |
| 8       | 2       | 10      | 5       |
| 9       | 6       | 7       | 2       |
| 11      | 5       | 7       | 3       |
| 13      | 7       | 11      | 7       |

18. Silver Springs Moving and Storage Inc. is studying the relationship between the number of rooms in a move and the number of labor hours required for the move. As part of the analysis, the CFO of Silver Springs developed the following scatter diagram.



- a. How many moves are in the sample?  
 b. Does it appear that more labor hours are required as the number of rooms increases, or do labor hours decrease as the number of rooms increases?
19. The Director of Planning for Devine Dining Inc. wishes to study the relationship between the gender of a guest and whether the guest orders dessert. To investigate the relationship, the manager collected the following information on 200 recent customers.

| Dessert Ordered | Gender    |           | Total      |
|-----------------|-----------|-----------|------------|
|                 | Male      | Female    |            |
| Yes             | 32        | 15        | 47         |
| No              | <u>68</u> | <u>85</u> | <u>153</u> |
| Total           | 100       | 100       | 200        |

- a. What is the level of measurement of the two variables?  
 b. What is the above table called?  
 c. Does the evidence in the table suggest men are more likely to order dessert than women? Explain why.
20. Ski Resorts of Vermont Inc. is considering a merger with Gulf Shores Beach Resorts Inc. of Alabama. The board of directors surveyed 50 stockholders concerning their position on the merger. The results are reported below.

| Number of Shares Held | Opinion  |           |           | Total     |
|-----------------------|----------|-----------|-----------|-----------|
|                       | Favor    | Oppose    | Undecided |           |
| Under 200             | 8        | 6         | 2         | 16        |
| 200 up to 1,000       | 6        | 8         | 1         | 15        |
| Over 1,000            | <u>6</u> | <u>12</u> | <u>1</u>  | <u>19</u> |
| Total                 | 20       | 26        | 4         | 50        |

- a. What level of measurement is used in this table?  
 b. What is this table called?  
 c. What group seems most strongly opposed to the merger?

## CHAPTER SUMMARY

- I. A dot plot shows the range of values on the horizontal axis and the number of observations for each value on the vertical axis.
  - A. Dot plots report the details of each observation.
  - B. They are useful for comparing two or more data sets.
- II. Measures of location also describe the shape of a set of observations.
  - A. Quartiles divide a set of observations into four equal parts.
    1. Twenty-five percent of the observations are less than the first quartile, 50% are less than the second quartile, and 75% are less than the third quartile.
    2. The interquartile range is the difference between the third quartile and the first quartile.
  - B. Deciles divide a set of observations into 10 equal parts and percentiles into 100 equal parts.
- III. A box plot is a graphic display of a set of data.
  - A. A box is drawn enclosing the regions between the first quartile and the third quartile.
    1. A line is drawn inside the box at the median value.
    2. Dotted line segments are drawn from the third quartile to the largest value to show the highest 25% of the values and from the first quartile to the smallest value to show the lowest 25% of the values.
  - B. A box plot is based on five statistics: the maximum and minimum values, the first and third quartiles, and the median.

IV. The coefficient of skewness is a measure of the symmetry of a distribution.

A. There are two formulas for the coefficient of skewness.

1. The formula developed by Pearson is:

$$sk = \frac{3(\bar{x} - \text{Median})}{s} \tag{4-2}$$

2. The coefficient of skewness computed by statistical software is:

$$sk = \frac{n}{(n-1)(n-2)} \left[ \sum \left( \frac{x - \bar{x}}{s} \right)^3 \right] \tag{4-3}$$

V. A scatter diagram is a graphic tool to portray the relationship between two variables.

A. Both variables are measured with interval or ratio scales.

B. If the scatter of points moves from the lower left to the upper right, the variables under consideration are directly or positively related.

C. If the scatter of points moves from the upper left to the lower right, the variables are inversely or negatively related.

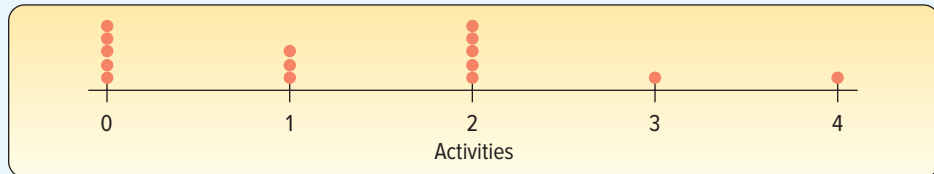
VI. A contingency table is used to classify nominal-scale observations according to two characteristics.

### PRONUNCIATION KEY

| SYMBOL | MEANING                | PRONUNCIATION |
|--------|------------------------|---------------|
| $L_p$  | Location of percentile | L sub p       |
| $Q_1$  | First quartile         | Q sub 1       |
| $Q_3$  | Third quartile         | Q sub 3       |

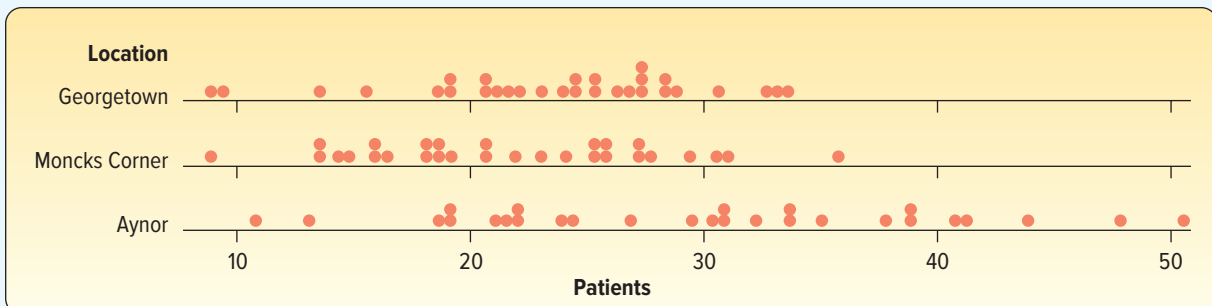
### CHAPTER EXERCISES

21. A sample of students attending Southeast Florida University is asked the number of social activities in which they participated last week. The chart below was prepared from the sample data.



- What is the name given to this chart?
- How many students were in the study?
- How many students reported attending no social activities?

22. Doctor's Care is a walk-in clinic, with locations in Georgetown, Moncks Corner, and Aynor, at which patients may receive treatment for minor injuries, colds, and flu, as well as physical examinations. The following charts report the number of patients treated in each of the three locations last month.



Describe the number of patients served at the three locations each day. What are the maximum and minimum numbers of patients served at each of the locations?

23. **FILE** In recent years, due to low interest rates, many homeowners refinanced their home mortgages. Linda Lahey is a mortgage officer at Down River Federal Savings and Loan. Below is the amount refinanced for 20 loans she processed last week. The data are reported in thousands of dollars and arranged from smallest to largest.

|       |       |      |      |      |      |      |      |      |
|-------|-------|------|------|------|------|------|------|------|
| 59.2  | 59.5  | 61.6 | 65.5 | 66.6 | 72.9 | 74.8 | 77.3 | 79.2 |
| 83.7  | 85.6  | 85.8 | 86.6 | 87.0 | 87.1 | 90.2 | 93.3 | 98.6 |
| 100.2 | 100.7 |      |      |      |      |      |      |      |

- a. Find the median, first quartile, and third quartile.  
 b. Find the 26th and 83rd percentiles.  
 c. Draw a box plot of the data.
24. **FILE** A study is made by the recording industry in the United States of the number of music CDs owned by 25 senior citizens and 30 young adults. The information is reported below.

| Seniors |     |     |     |     |     |     |     |     |     |
|---------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 28      | 35  | 41  | 48  | 52  | 81  | 97  | 98  | 98  | 99  |
| 118     | 132 | 133 | 140 | 145 | 147 | 153 | 158 | 162 | 174 |
| 177     | 180 | 180 | 187 | 188 |     |     |     |     |     |

| Young Adults |     |     |     |     |     |     |     |     |     |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 81           | 107 | 113 | 147 | 147 | 175 | 183 | 192 | 202 | 209 |
| 233          | 251 | 254 | 266 | 283 | 284 | 284 | 316 | 372 | 401 |
| 417          | 423 | 490 | 500 | 507 | 518 | 550 | 557 | 590 | 594 |

- a. Find the median and the first and third quartiles for the number of CDs owned by senior citizens. Develop a box plot for the information.  
 b. Find the median and the first and third quartiles for the number of CDs owned by young adults. Develop a box plot for the information.  
 c. Compare the number of CDs owned by the two groups.
25. **FILE** The corporate headquarters of *Bank.com*, an online banking company, is located in downtown Philadelphia. The director of human resources is making a study of the time it takes employees to get to work. The city is planning to offer incentives to each downtown employer if they will encourage their employees to use public transportation. Below is a listing of the time to get to work this morning according to whether the employee used public transportation or drove a car.

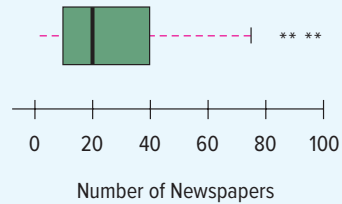
| Public Transportation |    |    |    |    |    |    |    |    |    |
|-----------------------|----|----|----|----|----|----|----|----|----|
| 23                    | 25 | 25 | 30 | 31 | 31 | 32 | 33 | 35 | 36 |
| 37                    | 42 |    |    |    |    |    |    |    |    |

| Private |    |    |    |    |    |    |    |    |    |
|---------|----|----|----|----|----|----|----|----|----|
| 32      | 32 | 33 | 34 | 37 | 37 | 38 | 38 | 38 | 39 |
| 40      | 44 |    |    |    |    |    |    |    |    |

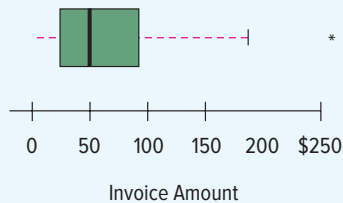
- a. Find the median and the first and third quartiles for the time it took employees using public transportation. Develop a box plot for the information.  
 b. Find the median and the first and third quartiles for the time it took employees who drove their own vehicle. Develop a box plot for the information.  
 c. Compare the times of the two groups.



26. The following box plot shows the number of daily newspapers published in each state and the District of Columbia. Write a brief report summarizing the number published. Be sure to include information on the values of the first and third quartiles, the median, and whether there is any skewness. If there are any outliers, estimate their value.



27. Walter Gogel Company is an industrial supplier of fasteners, tools, and springs. The amounts of its invoices vary widely, from less than \$20.00 to more than \$400.00. During the month of January, the company sent out 80 invoices. Here is a box plot of these invoices. Write a brief report summarizing the invoice amounts. Be sure to include information on the values of the first and third quartiles, the median, and whether there is any skewness. If there are any outliers, approximate the value of these invoices.



28. **FILE** The American Society of PeriAnesthesia Nurses (ASPAN; [www.aspan.org](http://www.aspan.org)) is a national organization serving nurses practicing in ambulatory surgery, preanesthesia, and postanesthesia care. The organization consists of the 40 components listed below.

| State/Region           | Membership | State/Region                                  | Membership |
|------------------------|------------|---|------------|
| Alabama                | 95         | New Jersey, Bermuda                           | 517        |
| Arizona                | 399        | Alaska, Idaho, Montana,<br>Oregon, Washington | 708        |
| Maryland, Delaware, DC | 531        | New York                                      | 891        |
| Connecticut            | 239        | Ohio  | 708        |
| Florida                | 631        | Oklahoma                                      | 171        |
| Georgia                | 384        | Arkansas                                      | 68         |
| Hawaii                 | 73         | California                                    | 1,165      |
| Illinois               | 562        | New Mexico                                    | 79         |
| Indiana                | 270        | Pennsylvania                                  | 575        |
| Iowa                   | 117        | Rhode Island                                  | 53         |
| Kentucky               | 197        | Colorado                                      | 409        |
| Louisiana              | 258        | South Carolina                                | 237        |
| Michigan               | 411        | Texas   | 1,026      |
| Massachusetts          | 480        | Tennessee                                     | 167        |
| Maine                  | 97         | Utah  | 67         |
| Minnesota, Dakotas     | 289        | Virginia                                      | 414        |
| Missouri, Kansas       | 282        | Vermont,<br>New Hampshire                     | 144        |
| Mississippi            | 90         | Wisconsin                                     | 311        |
| Nebraska               | 115        | West Virginia                                 | 62         |
| North Carolina         | 542        |   |            |
| Nevada                 | 106        |   |            |

Use statistical software to answer the following questions.

- a. Find the mean, median, and standard deviation of the number of members per component.
  - b. Find the coefficient of skewness, using the software. What do you conclude about the shape of the distribution of component size?
  - c. Compute the first and third quartiles using formula (4–1).
  - d. Develop a box plot. Are there any outliers? Which components are outliers? What are the limits for outliers?
29. **FILE** McGivern Jewelers is located in the Levis Square Mall just south of Toledo, Ohio. Recently it posted an advertisement on a social media site reporting the shape, size, price, and cut grade for 33 of its diamonds currently in stock. The information is reported below.

| Shape    | Size (carats) | Price    | Cut Grade       | Shape    | Size (carats) | Price   | Cut Grade       |
|----------|---------------|----------|-----------------|----------|---------------|---------|-----------------|
| Princess | 5.03          | \$44,312 | Ideal cut       | Round    | 0.77          | \$2,828 | Ultra ideal cut |
| Round    | 2.35          | 20,413   | Premium cut     | Oval     | 0.76          | 3,808   | Premium cut     |
| Round    | 2.03          | 13,080   | Ideal cut       | Princess | 0.71          | 2,327   | Premium cut     |
| Round    | 1.56          | 13,925   | Ideal cut       | Marquise | 0.71          | 2,732   | Good cut        |
| Round    | 1.21          | 7,382    | Ultra ideal cut | Round    | 0.70          | 1,915   | Premium cut     |
| Round    | 1.21          | 5,154    | Average cut     | Round    | 0.66          | 1,885   | Premium cut     |
| Round    | 1.19          | 5,339    | Premium cut     | Round    | 0.62          | 1,397   | Good cut        |
| Emerald  | 1.16          | 5,161    | Ideal cut       | Round    | 0.52          | 2,555   | Premium cut     |
| Round    | 1.08          | 8,775    | Ultra ideal cut | Princess | 0.51          | 1,337   | Ideal cut       |
| Round    | 1.02          | 4,282    | Premium cut     | Round    | 0.51          | 1,558   | Premium cut     |
| Round    | 1.02          | 6,943    | Ideal cut       | Round    | 0.45          | 1,191   | Premium cut     |
| Marquise | 1.01          | 7,038    | Good cut        | Princess | 0.44          | 1,319   | Average cut     |
| Princess | 1.00          | 4,868    | Premium cut     | Marquise | 0.44          | 1,319   | Premium cut     |
| Round    | 0.91          | 5,106    | Premium cut     | Round    | 0.40          | 1,133   | Premium cut     |
| Round    | 0.90          | 3,921    | Good cut        | Round    | 0.35          | 1,354   | Good cut        |
| Round    | 0.90          | 3,733    | Premium cut     | Round    | 0.32          | 896     | Premium cut     |
| Round    | 0.84          | 2,621    | Premium cut     |          |               |         |                 |

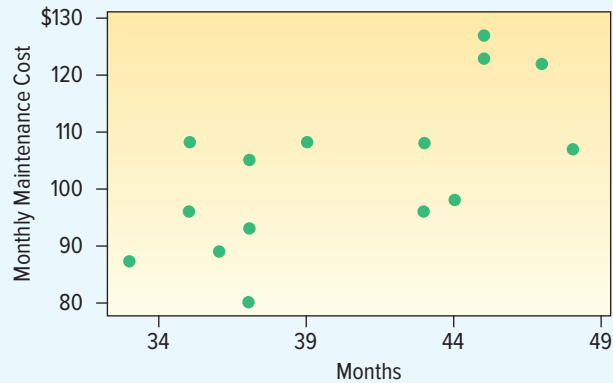
- a. Develop a box plot of the variable price and comment on the result. Are there any outliers? What is the median price? What are the values of the first and the third quartiles?
  - b. Develop a box plot of the variable size and comment on the result. Are there any outliers? What is the median price? What are the values of the first and the third quartiles?
  - c. Develop a scatter diagram between the variables price and size. Be sure to put price on the vertical axis and size on the horizontal axis. Does there seem to be an association between the two variables? Is the association direct or indirect? Does any point seem to be different from the others?
  - d. Develop a contingency table for the variables shape and cut grade. What is the most common cut grade? What is the most common shape? What is the most common combination of cut grade and shape?
30. **FILE** Listed below is the amount of commissions earned last month for the eight members of the sales staff at Best Electronics. Calculate the coefficient of skewness using both methods. Hint: Use of a spreadsheet will expedite the calculations.

980.9 1,036.5 1,099.5 1,153.9 1,409.0 1,456.4 1,718.4 1,721.2

31. **FILE** Listed below is the number of car thefts in a large city over the last week. Calculate the coefficient of skewness using both methods. Hint: Use of a spreadsheet will expedite the calculations.

3 12 13 7 8 3 8

32. The manager of Information Services at Wilkin Investigations, a private investigation firm, is studying the relationship between the age (in months) of a combination printer, copier, and fax machine and its monthly maintenance cost. For a sample of 15 machines, the manager developed the following chart. What can the manager conclude about the relationship between the variables?



33. **FILE** An auto insurance company reported the following information regarding the age of a driver and the number of accidents reported last year. Develop a scatter diagram for the data and write a brief summary.

| Age | Accidents | Age | Accidents |
|-----|-----------|-----|-----------|
| 16  | 4         | 23  | 0         |
| 24  | 2         | 27  | 1         |
| 18  | 5         | 32  | 1         |
| 17  | 4         | 22  | 3         |

34. Wendy's offers eight different condiments (mustard, ketchup, onion, mayonnaise, pickle, lettuce, tomato, and relish) on hamburgers. A store manager collected the following information on the number of condiments ordered and the age group of the customer. What can you conclude regarding the information? Who tends to order the most or least number of condiments?

| Number of Condiments | Age      |             |             |             |
|----------------------|----------|-------------|-------------|-------------|
|                      | Under 18 | 18 up to 40 | 40 up to 60 | 60 or Older |
| 0                    | 12       | 18          | 24          | 52          |
| 1                    | 21       | 76          | 50          | 30          |
| 2                    | 39       | 52          | 40          | 12          |
| 3 or more            | 71       | 87          | 47          | 28          |

35. Here is a table showing the number of employed and unemployed workers 20 years or older by gender in the United States.

| Gender | Number of Workers (000) |            |
|--------|-------------------------|------------|
|        | Employed                | Unemployed |
| Men    | 70,415                  | 4,209      |
| Women  | 61,402                  | 3,314      |

- How many workers were studied?
- What percent of the workers were unemployed?
- Compare the percent unemployed for the men and the women.

## DATA ANALYTICS

(The data for these exercises are available in Connect.)

- 36. FILE** Refer to the North Valley real estate data recorded on homes sold during the last year. Prepare a report on the selling prices of the homes based on the answers to the following questions.
- Compute the minimum, maximum, median, and the first and the third quartiles of price. Create a box plot. Comment on the distribution of home prices.
  - Develop a scatter diagram with price on the vertical axis and the size of the home on the horizontal. Is there a relationship between these variables? Is the relationship direct or indirect?
  - For homes without a pool, develop a scatter diagram with price on the vertical axis and the size of the home on the horizontal. Do the same for homes with a pool. How do the relationships between price and size for homes without a pool and homes with a pool compare?
- 37. FILE** Refer to the Baseball 2016 data that report information on the 30 Major League Baseball teams for the 2016 season.
- In the data set, the year opened is the first year of operation for that stadium. For each team, use this variable to create a new variable, stadium age, by subtracting the value of the variable year opened from the current year. Develop a box plot with the new variable, stadium age. Are there any outliers? If so, which of the stadiums are outliers?
  - Using the variable salary create a box plot. Are there any outliers? Compute the quartiles using formula  $(4-1)$ . Write a brief summary of your analysis.
  - Draw a scatter diagram with the variable wins on the vertical axis and salary on the horizontal axis. What are your conclusions?
  - Using the variable wins draw a dot plot. What can you conclude from this plot?
- 38. FILE** Refer to the Lincolnville School District bus data.
- Referring to the maintenance cost variable, develop a box plot. What are the minimum, first quartile, median, third quartile, and maximum values? Are there any outliers?
  - Using the median maintenance cost, develop a contingency table with bus manufacturer as one variable and whether the maintenance cost was above or below the median as the other variable. What are your conclusions?

## PRACTICE TEST

### Part 1—Objective

- A graph for displaying data in which each individual value is represented along a number line is called a \_\_\_\_\_.
- A \_\_\_\_\_ is a graphical display based on five statistics: the maximum and minimum values, the first and third quartiles, and the median.
- A \_\_\_\_\_ is a graphical technique used to show the relationship between two interval- or ratio-scaled variables.
- A \_\_\_\_\_ table is used to classify observations according to two identifiable characteristics.
- \_\_\_\_\_ divide a set of observations into four equal parts.
- \_\_\_\_\_ divide a set of observations into 100 equal parts.
- The coefficient of \_\_\_\_\_ measures the symmetry of a distribution.
- The \_\_\_\_\_ is the point below which one-fourth of the ranked values lie.
- The \_\_\_\_\_ is the difference between the first and third quartiles.

### Part 2—Problems

- Eleven insurance companies reported their market capitalization (in millions of dollars) for the most recent fiscal year as:

15   17   23   26   27   35   72   88   91   98   102

- Draw a dot plot of the data.
- Determine the median market capitalization.

- c. Compute the first quartile of market capitalization.
  - d. Find the 75th percentile of market capitalization.
  - e. Make a box plot of the data.
2. A Texas farm co-op sponsored a health screening for its members. Part of the process included a blood pressure screen. The results of the blood pressure screen are summarized by age groups in the following table:

| Blood Pressure | Age       |             |           | Total      |
|----------------|-----------|-------------|-----------|------------|
|                | Under 30  | 30 up to 60 | Over 60   |            |
| Low            | 21        | 29          | 37        | 87         |
| Medium         | 45        | 82          | 91        | 218        |
| High           | <u>23</u> | <u>46</u>   | <u>75</u> | <u>144</u> |
| Total          | 89        | 157         | 203       | 449        |

- a. What fraction of the members have high blood pressure?
- b. What fraction of the “Under 30” members have low blood pressure?
- c. Is there a relationship between age and blood pressure? Describe it.

# A Survey of Probability Concepts

# 5



©Yuri Yavnik/Shutterstock

- ▲ **RECENT SURVEYS** indicate 60% of tourists to China visited the Forbidden City, the Temple of Heaven, the Great Wall, and other historical sites in or near Beijing. Forty percent visited Xi'an and its magnificent terra-cotta soldiers, horses, and chariots, which lay buried for over 2,000 years. Thirty percent of the tourists went to both Beijing and Xi'an. What is the probability that a tourist visited at least one of these places? (See Exercise 68 and **LO5-3**.)

## LEARNING OBJECTIVES

---

*When you have completed this chapter, you will be able to:*

- LO5-1** Define the terms *probability*, *experiment*, *event*, and *outcome*.
- LO5-2** Assign probabilities using a classical, empirical, or subjective approach.
- LO5-3** Calculate probabilities using the rules of addition.
- LO5-4** Calculate probabilities using the rules of multiplication.
- LO5-5** Compute probabilities using a contingency table.
- LO5-6** Determine the number of outcomes using principles of counting.

## INTRODUCTION

The emphasis in Chapters 2, 3, and 4 is on descriptive statistics. In Chapter 2, we organize the profits on 180 vehicles sold by the Applewood Auto Group into a frequency distribution. This frequency distribution shows the smallest and the largest profits and where the largest concentration of data occurs. In Chapter 3, we use numerical measures of location and dispersion to locate a typical profit on vehicle sales and to examine the variation in the profit of a sale. We describe the variation in the profits with such measures of dispersion as the range and the standard deviation. In Chapter 4, we develop charts and graphs, such as a scatter diagram or a dot plot, to further describe the data graphically.

Descriptive statistics is concerned with summarizing data collected from past events. We now turn to the second facet of statistics, namely, *computing the chance that something will occur in the future*. This facet of statistics is called **statistical inference** or **inferential statistics**.

Seldom does a decision maker have complete information to make a decision. For example:



©BallDa/Shutterstock

- Toys and Things, a toy and puzzle manufacturer, recently developed a new game based on sports trivia. It wants to know whether sports buffs will purchase the game. “Slam Dunk” and “Home Run” are two of the names under consideration. To investigate, the president of Toys and Things decided to hire a market research firm. The firm selected a sample of 800 consumers from the population and asked each respondent for a reaction to the new game and its proposed titles. Using the sample results, the company can estimate the proportion of the population that will purchase the game.

### STATISTICS IN ACTION

Government statistics show there are about 1.7 automobile-caused fatalities for every 100,000,000 vehicle-miles. If you drive 1 mile to the store to buy your lottery ticket and then return home, you have driven 2 miles. Thus the probability that you will join this statistical group on your next 2-mile round trip is  $2 \times 1.7/100,000,000 = 0.000000034$ . This can also be stated as “One in 29,411,765.” Thus, if you drive to the store to buy your Powerball ticket, your chance of being killed (or killing someone else) is more than 4 times greater than the chance that you will win the Powerball jackpot, one chance in 120,526,770.

<http://www.durangobill.com/PowerballOdds.html>

- The quality assurance department of a U.S. Steel mill must assure management that the quarter-inch wire being produced has an acceptable tensile strength. Clearly, not all the wire produced can be tested for tensile strength because testing requires the wire to be stretched until it breaks—thus destroying it. So a random sample of 10 pieces is selected and tested. Based on the test results, all the wire produced is deemed to be either acceptable or unacceptable.
- Other questions involving uncertainty are: Should the daytime drama *Days of Our Lives* be discontinued immediately? Will a newly developed mint-flavored cereal be profitable if marketed? Will Charles Linden be elected to county auditor in Batavia County?

Statistical inference deals with conclusions about a population based on a sample taken from that population. (The populations for the preceding illustrations are all consumers who like sports trivia games, all the quarter-inch steel wire produced, all television viewers who watch soaps, all who purchase breakfast cereal, and so on.)

Because there is uncertainty in decision making, it is important that all the known risks involved be scientifically evaluated. Helpful in this evaluation is *probability theory*, often referred to as the science of uncertainty. Probability theory allows the decision maker to analyze the risks and minimize the gamble inherent, for example, in marketing a new product or accepting an incoming shipment possibly containing defective parts.

Because probability concepts are so important in the field of statistical inference (to be discussed starting with Chapter 8), this chapter introduces the basic language of probability, including such terms as *experiment*, *event*, *subjective probability*, and *addition* and *multiplication rules*.

**LO5-1**

Define the terms *probability*, *experiment*, *event*, and *outcome*.

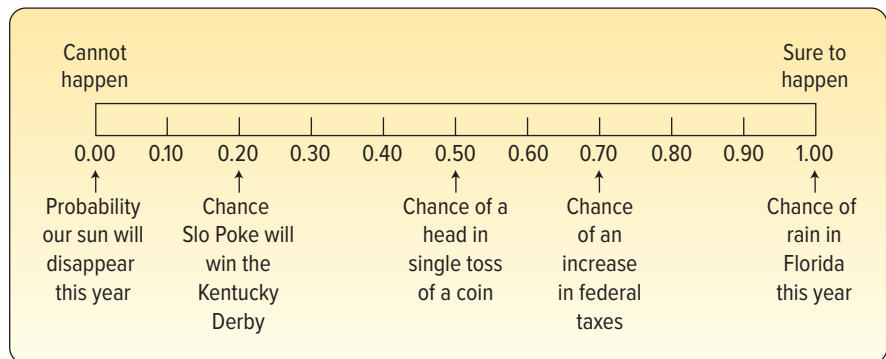
## WHAT IS A PROBABILITY?

No doubt you are familiar with terms such as *probability*, *chance*, and *likelihood*. They are often used interchangeably. The weather forecaster announces that there is a 70% chance of rain for Super Bowl Sunday. Based on a survey of consumers who tested a newly developed toothpaste with a banana flavor, the probability is .03 that, if marketed, it will be a financial success. (This means that the chance of the banana-flavor toothpaste being accepted by the public is rather remote.) What is a **probability**? In general, it is a numerical value that describes the chance that something will happen.

**PROBABILITY** A value between zero and one, inclusive, describing the relative possibility (chance or likelihood) an event will occur.

A probability is frequently expressed as a decimal, such as .70, .27, or .50, or a percent, such as 70%, 27%, or 50%. It also may be reported as a fraction, such as  $7/10$ ,  $27/100$ , or  $1/2$ . It can assume any number from 0 to 1, inclusive. Expressed as a percentage, the range is between 0% and 100%, inclusive. If a company has only five sales regions, and each region's name or number is written on a slip of paper and the slips put in a hat, the probability of selecting one of the five regions is  $1/5$ . The probability of selecting from the hat a slip of paper that reads "Pittsburgh Steelers" is 0. Thus, the probability of 1 represents something that is certain to happen, and the probability of 0 represents something that cannot happen.

The closer a probability is to 0, the more improbable it is the event will happen. The closer the probability is to 1, the more likely it will happen. The relationship is shown in the following diagram along with a few of our personal beliefs. You might, however, select a different probability for Slo Poke's chances to win the Kentucky Derby or for an increase in federal taxes.



Sometimes, the likelihood of an event is expressed using the term *odds*. To explain, someone says the odds are "five to two" that an event will occur. This means that in a total of seven trials ( $5 + 2$ ), the event will occur five times and not occur two times. Using odds, we can compute the probability that the event occurs as  $5/(5 + 2)$  or  $5/7$ . So, if the odds in favor of an event are  $x$  to  $y$ , the probability of the event is  $x/(x + y)$ .

Three key words are used in the study of probability: **experiment**, **outcome**, and **event**. These terms are used in our everyday language, but in statistics they have specific meanings.

**EXPERIMENT** A process that leads to the occurrence of one and only one of several possible results.



This definition is more general than the one used in the physical sciences, where we picture someone manipulating test tubes or microscopes. In reference to probability, an experiment has two or more possible results, and it is uncertain which will occur.



**OUTCOME** A particular result of an experiment.

For example, the tossing of a coin is an experiment. You are unsure of the outcome. When a coin is tossed, one particular outcome is a “head.” The alternative outcome is a “tail.” Similarly, asking 500 college students if they would travel more than 100 miles to attend a Mumford and Sons concert is an experiment. In this experiment, one possible outcome is that 273 students indicate they would travel more than 100 miles to attend the concert. Another outcome is that 317 students would attend the concert. Still another outcome is that 423 students indicate they would attend the concert. When one or more of the experiment’s outcomes are observed, we call this an event.

**EVENT** A collection of one or more outcomes of an experiment.

Examples to clarify the definitions of the terms *experiment*, *outcome*, and *event* are presented in the following figure.

In the die-rolling experiment, there are six possible outcomes, but there are many possible events. When counting the number of members of the board of directors for Fortune 500 companies over 60 years of age, the number of possible outcomes can be anywhere from zero to the total number of members. There are an even larger number of possible events in this experiment.

|                       |   |   |
|-----------------------|---|---|
|                       |      |                           |
| Experiment            | Roll a die  | Count the number of members of the board of directors for Fortune 500 companies who are over 60 years of age  |
| All possible outcomes | Observe a 1<br>Observe a 2<br>Observe a 3<br>Observe a 4<br>Observe a 5<br>Observe a 6  | None is over 60<br>One is over 60<br>Two are over 60<br>...<br>29 are over 60<br>...<br>48 are over 60<br>... |
| Some possible events  | Observe an even number<br>Observe a number greater than 4<br>Observe a number 3 or less | More than 13 are over 60<br>Fewer than 20 are over 60   |

**SELF-REVIEW 5-1**



RedLine Productions recently developed a new video game. Its playability is to be tested by 80 veteran game players.

- (a) What is the experiment?
- (b) What is one possible outcome?
- (c) Suppose 65 of the 80 players testing the new game said they liked it. Is 65 a probability?
- (d) The probability that the new game will be a success is computed to be  $-1.0$ . Comment.
- (e) Specify one possible event.

**LO5-2**

Assign probabilities using a classical, empirical, or subjective approach.

**APPROACHES TO ASSIGNING PROBABILITIES**

There are three ways to assign a probability to an event: classical, empirical, and subjective. The classical and empirical methods are objective and are based on information and data. The subjective method is based on a person’s belief or estimate of an event’s likelihood.

**Classical Probability**

**Classical probability** is based on the assumption that the outcomes of an experiment are *equally likely*. Using the classical viewpoint, the probability of an event happening is computed by dividing the number of favorable outcomes by the number of possible outcomes:

**CLASSICAL PROBABILITY**      Probability of an event =  $\frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$       **(5-1)**

**EXAMPLE**

Consider an experiment of rolling a six-sided die. What is the probability of the event “an even number of spots appear face up”?

**SOLUTION**

The possible outcomes are:

|              |             |
|--------------|-------------|
| a one-spot   | a four-spot |
| a two-spot   | a five-spot |
| a three-spot | a six-spot  |

There are three “favorable” outcomes (a two, a four, and a six) in the collection of six equally likely possible outcomes. Therefore:

$$\begin{aligned} \text{Probability of an even number} &= \frac{3}{6} && \leftarrow \boxed{\frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}} \\ &= .5 && \leftarrow \end{aligned}$$

The **mutually exclusive** concept appeared earlier in our study of frequency distributions in Chapter 2. Recall that we create classes so that a particular value is included in only one of the classes and there is no overlap between classes. Thus, only one of several events can occur at a particular time.

**MUTUALLY EXCLUSIVE** The occurrence of one event means that none of the other events can occur at the same time.

A decision to attend a four-year university presents mutually exclusive outcomes. A high school senior decides either to attend or not. A decision to do both is not logical. A manufactured part is acceptable or unacceptable. The part cannot be both acceptable and unacceptable at the same time. In a sample of manufactured parts, the event of selecting an unacceptable part and the event of selecting an acceptable part are mutually exclusive.

If an experiment has a set of events that includes every possible outcome, such as the events “an even number” and “an odd number” in the die-tossing experiment, then the set of events is **collectively exhaustive**. For the die-tossing experiment, every outcome will be either even or odd. So the set is collectively exhaustive.

**COLLECTIVELY EXHAUSTIVE** At least one of the events must occur when an experiment is conducted.

If the set of events is collectively exhaustive and the events are mutually exclusive, the sum of the probabilities is 1. Historically, the classical approach to probability was developed and applied in the 17th and 18th centuries to games of chance, such as cards and dice. It is unnecessary to do an experiment to determine the probability of an event occurring using the classical approach because the total number of outcomes is known before the experiment. The flip of a coin has two possible outcomes; the roll of a die has six possible outcomes. We can logically arrive at the probability of getting a tail on the toss of one coin or three heads on the toss of three coins.

The classical approach to probability can also be applied to lotteries. In South Carolina, one of the games of the Education Lottery is “Pick 3.” A person buys a lottery ticket and selects three numbers between 0 and 9. Once per week, the three numbers are randomly selected from a machine that tumbles three containers each with balls numbered 0 through 9. One way to win is to match the numbers and the order of the numbers. Given that 1,000 possible outcomes exist (000 through 999), the probability of winning with any three-digit number is 0.001, or 1 in 1,000.

## Empirical Probability

**Empirical or relative frequency** is the second type of objective probability. It is based on the number of times an event occurs as a proportion of a known number of trials.

**EMPIRICAL PROBABILITY** The probability of an event happening is the fraction of the time similar events happened in the past.

The formula to determine an empirical probability is:

$$\text{Empirical probability} = \frac{\text{Number of times the event occurs}}{\text{Total number of observations}}$$

The empirical approach to probability is based on what is called the **law of large numbers**. The key to establishing probabilities empirically is that more observations will provide a more accurate estimate of the probability.

**LAW OF LARGE NUMBERS** Over a large number of trials, the empirical probability of an event will approach its true probability.

To explain the law of large numbers, suppose we toss a fair coin. The result of each toss is either a head or a tail. With just one toss of the coin, the empirical probability for

heads is either zero or one. If we toss the coin a great number of times, the probability of the outcome of heads will approach .5. The following table reports the results of seven different experiments of flipping a fair coin 1, 10, 50, 100, 500, 1,000, and 10,000 times and then computing the relative frequency of heads. Note as we increase the number of trials, the empirical probability of a head appearing approaches .5, which is its value based on the classical approach to probability.

| Number of Trials | Number of Heads | Relative Frequency of Heads |
|------------------|-----------------|-----------------------------|
| 1                | 0               | .00                         |
| 10               | 3               | .30                         |
| 50               | 26              | .52                         |
| 100              | 52              | .52                         |
| 500              | 236             | .472                        |
| 1,000            | 494             | .494                        |
| 10,000           | 5,027           | .5027                       |

What have we demonstrated? Based on the classical definition of probability, the likelihood of obtaining a head in a single toss of a fair coin is .5. Based on the empirical or relative frequency approach to probability, the probability of the event happening approaches the same value based on the classical definition of probability.

This reasoning allows us to use the empirical or relative frequency approach to finding a probability. Here are some examples.

- Last semester, 80 students registered for Business Statistics 101 at Scandia University. Twelve students earned an A. Based on this information and the empirical approach to assigning a probability, we estimate the likelihood a student at Scandia will earn an A is .15.
- Stephen Curry of the Golden State Warriors made 363 out of 400 free throw attempts during the 2015–16 NBA season. Based on the empirical approach to probability, the likelihood of him making his next free throw attempt is .908.

Life insurance companies rely on past data to determine the acceptability of an applicant as well as the premium to be charged. Mortality tables list the likelihood a person of a particular age will die within the upcoming year. For example, the likelihood a 20-year-old female will die within the next year is .00105.

The empirical concept is illustrated with the following example.

### EXAMPLE

On February 1, 2003, the Space Shuttle *Columbia* exploded. This was the second disaster in 113 space missions for NASA. On the basis of this information, what is the probability that a future mission is successfully completed?

### SOLUTION

We use letters or numbers to simplify the equations.  $P$  stands for probability and  $A$  represents the event of a successful mission. In this case,  $P(A)$  stands for the probability a future mission is successfully completed.

$$\text{Probability of a successful flight} = \frac{\text{Number of successful flights}}{\text{Total number of flights}}$$

$$P(A) = \frac{11}{113} = .0973$$

We can use this as an estimate of probability. In other words, based on past experience, the probability is .0973 that a future space shuttle mission will be safely completed.

## Subjective Probability

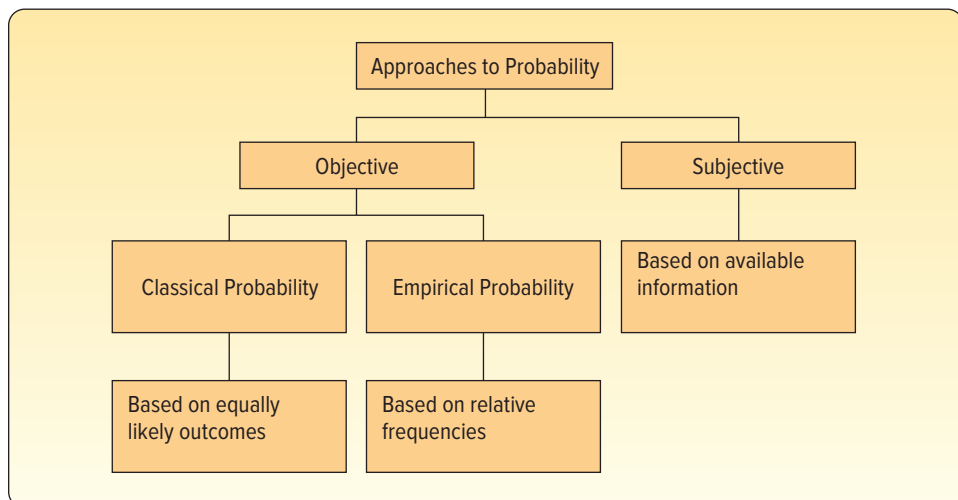
If there is little or no experience or information on which to base a probability, it is estimated subjectively. Essentially, this means an individual evaluates the available opinions and information and then estimates or assigns the probability. This probability is called a **subjective probability**.

**SUBJECTIVE CONCEPT OF PROBABILITY** The likelihood (probability) of a particular event happening that is assigned by an individual based on whatever information is available.

Illustrations of subjective probability are:

1. Estimating the likelihood the New England Patriots will play in the Super Bowl next year.
2. Estimating the likelihood you will be involved in an automobile accident during the next 12 months.
3. Estimating the likelihood the U.S. budget deficit will be reduced by half in the next 10 years.

The types of probability are summarized in Chart 5–1. A probability statement always assigns a likelihood to an event that has not yet occurred. There is, of course, considerable latitude in the degree of uncertainty that surrounds this probability, based primarily on the knowledge possessed by the individual concerning the underlying process. The individual possesses a great deal of knowledge about the toss of a die and can state that the probability that a one-spot will appear face up on the toss of a true die is one-sixth. But we know very little concerning the acceptance in the marketplace of a new and untested product. For example, even though a market research director tests a newly developed product in 40 retail stores and states that there is a 70% chance that the product will have sales of more than 1 million units, she has limited knowledge of how consumers will react when it is marketed nationally. In both cases (the case of the person rolling a die and the testing of a new product), the individual is assigning a probability value to an event of interest, and a difference exists only in the predictor's confidence in the precision of the estimate. However, regardless of the viewpoint, the same laws of probability (presented in the following sections) will be applied.



**CHART 5–1** Summary of Approaches to Probability

## SELF-REVIEW 5-2



1. One card will be randomly selected from a standard 52-card deck. What is the probability the card will be a queen? Which approach to probability did you use to answer this question?
2. The Center for Child Care reports on 539 children and the marital status of their parents. There are 333 married, 182 divorced, and 24 widowed parents. What is the probability a particular child chosen at random will have a parent who is divorced? Which approach did you use?
3. What is the probability you will save one million dollars by the time you retire? Which approach to probability did you use to answer this question?

## EXERCISES

1. Some people are in favor of reducing federal taxes to increase consumer spending, and others are against it. Two persons are selected and their opinions are recorded. Assuming no one is undecided, list the possible outcomes.
2. A quality control inspector selects a part to be tested. The part is then declared acceptable, repairable, or scrapped. Then another part is tested. List the possible outcomes of this experiment regarding two parts.
3. **FILE** A survey of 34 students at the Wall College of Business showed the following majors:

|            |    |
|------------|----|
| Accounting | 10 |
| Finance    | 5  |
| Economics  | 3  |
| Management | 6  |
| Marketing  | 10 |

- From the 34 students, suppose you randomly select a student.
- a. What is the probability he or she is a management major?
  - b. Which concept of probability did you use to make this estimate?
4. A large company must hire a new president. The board of directors prepares a list of five candidates, all of whom are equally qualified. Two of these candidates are members of a minority group. To avoid bias in the selection of the candidate, the company decides to select the president by lottery.
    - a. What is the probability one of the minority candidates is hired?
    - b. Which concept of probability did you use to make this estimate?
  5. In each of the following cases, indicate whether classical, empirical, or subjective probability is used.
    - a. A baseball player gets a hit in 30 out of 100 times at bat. The probability is .3 that he gets a hit in his next at bat.
    - b. A seven-member committee of students is formed to study environmental issues. What is the likelihood that any one of the seven is randomly chosen as the spokesperson?
    - c. You purchase a ticket for the Lotto Canada lottery. Over 5 million tickets were sold. What is the likelihood you will win the \$1 million jackpot?
    - d. The probability of an earthquake in northern California in the next 10 years above 5.0 on the Richter scale is .80.
  6. A firm will promote two employees out of a group of six men and three women.
    - a. List all possible outcomes.
    - b. What probability concept would be used to assign probabilities to the outcomes?
  7. A sample of 40 oil industry executives was selected to test a questionnaire. One question about environmental issues required a yes or no answer.
    - a. What is the experiment?
    - b. List one possible event.
    - c. Ten of the 40 executives responded yes. Based on these sample responses, what is the probability that an oil industry executive will respond yes?
    - d. What concept of probability does this illustrate?
    - e. Are each of the possible outcomes equally likely and mutually exclusive?

8. **FILE** A sample of 2,000 licensed drivers revealed the following number of speeding violations.

| Number of Violations | Number of Drivers |
|----------------------|-------------------|
| 0                    | 1,910             |
| 1                    | 46                |
| 2                    | 18                |
| 3                    | 12                |
| 4                    | 9                 |
| 5 or more            | 5                 |
| Total                | 2,000             |

- What is the experiment?
  - List one possible event.
  - What is the probability that a particular driver had exactly two speeding violations?
  - What concept of probability does this illustrate?
9. Bank of America customers select their own four-digit personal identification number (PIN) for use at ATMs.
- Think of this as an experiment and list four possible outcomes.
  - What is the probability that a customer will pick 2591 as their PIN?
  - Which concept of probability did you use to answer (b)?
10. An investor buys 100 shares of AT&T stock and records its price change daily.
- List several possible events for this experiment.
  - Which concept of probability did you use in (a)?

### LO5-3

Calculate probabilities using the rules of addition.

## RULES OF ADDITION FOR COMPUTING PROBABILITIES

There are two rules of addition, the special rule of addition and the general rule of addition. We begin with the special rule of addition.

### Special Rule of Addition

When we use the **special rule of addition**, the events must be *mutually exclusive*. Recall that mutually exclusive means that when one event occurs, none of the other events can occur at the same time. An illustration of mutually exclusive events in the die-tossing experiment is the events “a number 4 or larger” and “a number 2 or smaller.” If the outcome is in the first group {4, 5, and 6}, then it cannot also be in the second group {1 and 2}. Another illustration is a product coming off the assembly line cannot be defective and satisfactory at the same time.

If two events  $A$  and  $B$  are mutually exclusive, the special rule of addition states that the probability of one or the other event’s occurring equals the sum of their probabilities. This rule is expressed in the following formula:

#### SPECIAL RULE OF ADDITION

$$P(A \text{ or } B) = P(A) + P(B)$$

(5-2)

For three mutually exclusive events designated  $A$ ,  $B$ , and  $C$ , the rule is written:

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$$

An example will show the details.

### EXAMPLE



©BravissimoS/Shutterstock

A machine fills plastic bags with a mixture of beans, broccoli, and other vegetables. Most of the bags contain the correct weight, but because of the variation in the size of the beans and other vegetables, a package might be underweight or overweight. A check of 4,000 packages filled in the past month revealed:

| Weight       | Event | Number of Packages | Probability of Occurrence |
|--------------|-------|--------------------|---------------------------|
| Underweight  | $A$   | 100                | .025                      |
| Satisfactory | $B$   | 3,600              | .900                      |
| Overweight   | $C$   | 300                | .075                      |
|              |       | 4,000              | 1.000                     |

What is the probability that a particular package will be either underweight or overweight?

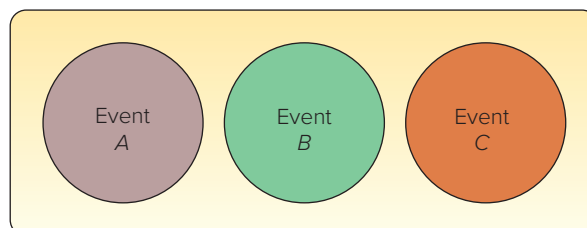
### SOLUTION

The outcome “underweight” is the event  $A$ . The outcome “overweight” is the event  $C$ . Applying the special rule of addition:

$$P(A \text{ or } C) = P(A) + P(C) = .025 + .075 = .10$$

Note that the events are mutually exclusive, meaning that a package of mixed vegetables cannot be underweight, satisfactory, and overweight at the same time. They are also collectively exhaustive; that is, a selected package must be either underweight, satisfactory, or overweight.

English logician J. Venn (1834–1923) developed a diagram to portray graphically the outcome of an experiment. The *mutually exclusive* concept and various other rules for combining probabilities can be illustrated using this device. To construct a Venn diagram, a space is first enclosed representing the total of all possible outcomes. This space is usually in the form of a rectangle. An event is then represented by a circular area that is drawn inside the rectangle proportional to the probability of the event. The following Venn diagram represents the *mutually exclusive* concept. There is no overlapping of events, meaning that the events are mutually exclusive. In the following Venn diagram, assume the events  $A$ ,  $B$ , and  $C$  are about equally likely.





## Complement Rule

The probability that a bag of mixed vegetables selected is underweight,  $P(A)$ , plus the probability that it is not an underweight bag, written  $P(\sim A)$  and read “not  $A$ ,” must logically equal 1. This is written:

$$P(A) + P(\sim A) = 1$$

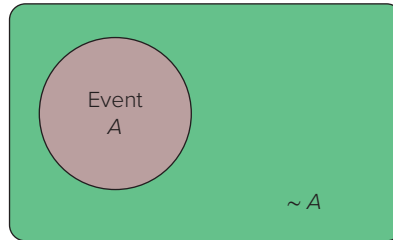
This can be revised to read:

### COMPLEMENT RULE

$$P(A) = 1 - P(\sim A)$$

(5–3)

This is the **complement rule**. It is used to determine the probability of an event occurring by subtracting the probability of the event not occurring from 1. This rule is useful because sometimes it is easier to calculate the probability of an event happening by determining the probability of it not happening and subtracting the result from 1. Notice that the events  $A$  and  $\sim A$  are mutually exclusive and collectively exhaustive. Therefore, the probabilities of  $A$  and  $\sim A$  sum to 1. A Venn diagram illustrating the complement rule is shown as:

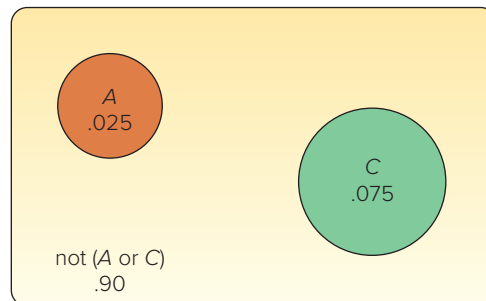


### EXAMPLE

Referring to the previous example/solution, the probability a bag of mixed vegetables is underweight is .025 and the probability of an overweight bag is .075. Use the complement rule to show the probability of a satisfactory bag is .900. Show the solution using a Venn diagram.

### SOLUTION

The probability the bag is unsatisfactory equals the probability the bag is overweight plus the probability it is underweight. That is,  $P(A \text{ or } C) = P(A) + P(C) = .025 + .075 = .100$ . The bag is satisfactory if it is not underweight or overweight, so  $P(B) = 1 - [P(A) + P(C)] = 1 - [.025 + .075] = 0.900$ . The Venn diagram portraying this situation is:



## SELF-REVIEW 5-3



A sample of employees of Worldwide Enterprises is to be surveyed about a new health care plan. The employees are classified as follows:

| Classification | Event | Number of Employees |
|----------------|-------|---------------------|
| Supervisors    | $A$   | 120                 |
| Maintenance    | $B$   | 50                  |
| Production     | $C$   | 1,460               |
| Management     | $D$   | 302                 |
| Secretarial    | $E$   | 68                  |

- (a) What is the probability that the first person selected is:
- either in maintenance or a secretary?
  - not in management?
- (b) Draw a Venn diagram illustrating your answers to part (a).
- (c) Are the events in part (a)(i) complementary or mutually exclusive or both?

### The General Rule of Addition

The outcomes of an experiment may not be mutually exclusive. For example, the Florida Tourist Commission selected a sample of 200 tourists who visited the state during the year. The survey revealed that 120 tourists went to Disney World and 100 went to Busch Gardens near Tampa. What is the probability that a person selected visited either Disney World or Busch Gardens? If the special rule of addition is used, the probability of selecting a tourist who went to Disney World is .60, found by  $120/200$ . Similarly, the probability of a tourist going to Busch Gardens is .50. The sum of these probabilities is 1.10. We know, however, that this probability cannot be greater than 1. The explanation is that many tourists visited both attractions and are being counted twice! A check of the survey responses revealed that 60 out of 200 sampled did, in fact, visit both attractions.

To answer our question, “What is the probability a selected person visited either Disney World or Busch Gardens?” (1) add the probability that a tourist visited Disney World and the probability he or she visited Busch Gardens, and (2) subtract the probability of visiting both. Thus:

$$\begin{aligned} P(\text{Disney or Busch}) &= P(\text{Disney}) + P(\text{Busch}) - P(\text{both Disney and Busch}) \\ &= .60 + .50 - .30 = .80 \end{aligned}$$

When two events both occur, the probability is called a **joint probability**. The probability (.30) that a tourist visits both attractions is an example of a joint probability.

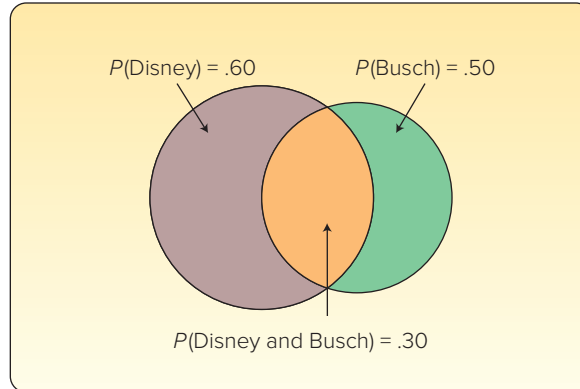


©Rostislav Glinsky/Shutterstock

The following Venn diagram shows two events that are not mutually exclusive. The two events overlap to illustrate the joint event that some people have visited both attractions.

#### STATISTICS IN ACTION

If you wish to get some attention at the next gathering you attend, announce that you believe that at least two people present were born on the same date—that is, the same day of the year but not necessarily the same year. If there are 30 people in the room, the probability of a duplicate is .706. If there are 60 people in the room, the probability is .994 that at least two people share the same birthday. With as few as 23 people the chances are even, that is .50, that at least two people share the same birthday. Hint: To compute this, find the probability everyone was born on a different day and use the complement rule. Try this in your class.



**JOINT PROBABILITY** A probability that measures the likelihood two or more events will happen concurrently.

So the general rule of addition, which is used to compute the probability of two events that are not mutually exclusive, is:

**GENERAL RULE OF ADDITION**  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$  (5-4)

For the expression  $P(A \text{ or } B)$ , the word *or* suggests that  $A$  may occur or  $B$  may occur. This also includes the possibility that  $A$  and  $B$  may occur. This use of *or* is sometimes called an **inclusive**. You could also write  $P(A \text{ or } B \text{ or both})$  to emphasize that the union of the events includes the intersection of  $A$  and  $B$ .

If we compare the general and special rules of addition, the important difference is determining if the events are mutually exclusive. If the events *are* mutually exclusive, then the joint probability  $P(A \text{ and } B)$  is 0 and we can use the special rule of addition. Otherwise, we must account for the joint probability and use the general rule of addition.

### EXAMPLE

What is the probability that a card chosen at random from a standard deck of cards will be either a king or a heart?

### SOLUTION

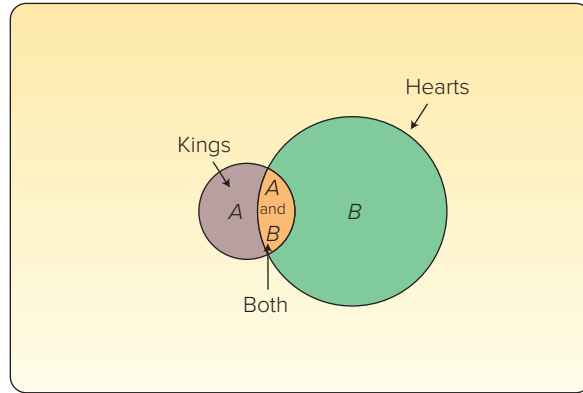
We may be inclined to add the probability of a king and the probability of a heart. But this creates a problem. If we do that, the king of hearts is counted with the kings and also with the hearts. So, if we simply add the probability of a king (there are 4 in a deck of 52 cards) to the probability of a heart (there are 13 in a deck of 52 cards) and report that 17 out of 52 cards meet the requirement, we have counted the king of hearts twice. We need to subtract 1 card from the 17 so the king of hearts is counted only once. Thus, there are 16 cards that are either hearts or kings. So the probability is  $16/52 = .3077$ .

| Card           | Probability                  | Explanation                            |
|----------------|------------------------------|--|
| King           | $P(A) = 4/52$                | 4 kings in a deck of 52 cards          |
| Heart          | $P(B) = 13/52$               | 13 hearts in a deck of 52 cards        |
| King of Hearts | $P(A \text{ and } B) = 1/52$ | 1 king of hearts in a deck of 52 cards |

From formula (5-4):

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) - P(A \text{ and } B) \\ &= 4/52 + 13/52 - 1/52 \\ &= 16/52, \text{ or } .3077 \end{aligned}$$

A Venn diagram portrays these outcomes, which are not mutually exclusive.



### SELF-REVIEW 5-4



Routine physical examinations are conducted annually as part of a health service program for General Concrete Inc. employees. It was discovered that 8% of the employees need corrective shoes, 15% need major dental work, and 3% need both corrective shoes and major dental work.

- (a) What is the probability that an employee selected at random will need either corrective shoes or major dental work?
- (b) Show this situation in the form of a Venn diagram.

### EXERCISES

11. The events  $A$  and  $B$  are mutually exclusive. Suppose  $P(A) = .30$  and  $P(B) = .20$ . What is the probability of either  $A$  or  $B$  occurring? What is the probability that neither  $A$  nor  $B$  will happen?
12. The events  $X$  and  $Y$  are mutually exclusive. Suppose  $P(X) = .05$  and  $P(Y) = .02$ . What is the probability of either  $X$  or  $Y$  occurring? What is the probability that neither  $X$  nor  $Y$  will happen?
13. **FILE** A study of 200 advertising firms revealed their income after taxes:

| Income after Taxes          | Number of Firms |
|-----------------------------|-----------------|
| Under \$1 million           | 102             |
| \$1 million to \$20 million | 61              |
| \$20 million or more        | 37              |

- a. What is the probability an advertising firm selected at random has under \$1 million in income after taxes?
- b. What is the probability an advertising firm selected at random has either an income between \$1 million and \$20 million, or an income of \$20 million or more? What rule of probability was applied?
14. The chair of the board of directors says, "There is a 50% chance this company will earn a profit, a 30% chance it will break even, and a 20% chance it will lose money next quarter."
  - a. Use an addition rule to find the probability the company will not lose money next quarter.
  - b. Use the complement rule to find the probability it will not lose money next quarter.
15. Suppose the probability you will get an A in this class is .25 and the probability you will get a B is .50. What is the probability your grade will be above a C?

16. Two coins are tossed. If  $A$  is the event “two heads” and  $B$  is the event “two tails,” are  $A$  and  $B$  mutually exclusive? Are they complements?
17. The probabilities of the events  $A$  and  $B$  are .20 and .30, respectively. The probability that both  $A$  and  $B$  occur is .15. What is the probability of either  $A$  or  $B$  occurring?
18. Let  $P(X) = .55$  and  $P(Y) = .35$ . Assume the probability that they both occur is .20. What is the probability of either  $X$  or  $Y$  occurring?
19. Suppose the two events  $A$  and  $B$  are mutually exclusive. What is the probability of their joint occurrence?
20. A student is taking two courses, history and math. The probability the student will pass the history course is .60, and the probability of passing the math course is .70. The probability of passing both is .50. What is the probability of passing at least one?
21. The aquarium at Sea Critters Depot contains 140 fish. Eighty of these fish are green swordtails (44 female and 36 male) and 60 are orange swordtails (36 female and 24 male). A fish is randomly captured from the aquarium:
  - a. What is the probability the selected fish is a green swordtail?
  - b. What is the probability the selected fish is male?
  - c. What is the probability the selected fish is a male green swordtail?
  - d. What is the probability the selected fish is either a male or a green swordtail?
22. A National Park Service survey of visitors to the Rocky Mountain region revealed that 50% visit Yellowstone Park, 40% visit the Tetons, and 35% visit both.
  - a. What is the probability a vacationer will visit at least one of these attractions?
  - b. What is the probability .35 called?
  - c. Are the events mutually exclusive? Explain.

**LO5-4**

Calculate probabilities using the rules of multiplication.

## RULES OF MULTIPLICATION TO CALCULATE PROBABILITY

In this section, we discuss the rules for computing the likelihood that two events both happen, or their joint probability. For example, 16% of the 2017 tax returns were prepared by H&R Block and 75% of those returns showed a refund. What is the likelihood a person’s tax form was prepared by H&R Block and the person received a refund? Venn diagrams illustrate this as the intersection of two events. To find the likelihood of two events happening, we use the rules of multiplication. There are two rules of multiplication: the special rule and the general rule.

### Special Rule of Multiplication

The special rule of multiplication requires that two events  $A$  and  $B$  are **independent**. Two events are independent if the occurrence of one event does not alter the probability of the occurrence of the other event.

**INDEPENDENCE** The occurrence of one event has no effect on the probability of the occurrence of another event.

One way to think about independence is to assume that events  $A$  and  $B$  occur at different times. For example, when event  $B$  occurs after event  $A$  occurs, does  $A$  have any effect on the likelihood that event  $B$  occurs? If the answer is no, then  $A$  and  $B$  are independent events. To illustrate independence, suppose two coins are tossed. The outcome of a coin toss (head or tail) is unaffected by the outcome of any other prior coin toss (head or tail).

For two independent events  $A$  and  $B$ , the probability that  $A$  and  $B$  will both occur is found by multiplying the two probabilities. This is the **special rule of multiplication** and is written symbolically as:

**SPECIAL RULE OF MULTIPLICATION**

$$P(A \text{ and } B) = P(A)P(B)$$

**(5–5)**

For three independent events,  $A$ ,  $B$ , and  $C$ , the special rule of multiplication used to determine the probability that all three events will occur is:

$$P(A \text{ and } B \text{ and } C) = P(A)P(B)P(C)$$

**EXAMPLE**

**STATISTICS IN ACTION**

In 2000 George W. Bush won the U.S. presidency by the slimmest of margins. Many election stories resulted, some involving voting irregularities, others raising interesting election questions. In a local Michigan election, there was a tie between two candidates for an elected position. To break the tie, the candidates drew a slip of paper from a box that contained two slips of paper, one marked "Winner" and the other unmarked. To determine which candidate drew first, election officials flipped a coin. The winner of the coin flip also drew the winning slip of paper. But was the coin flip really necessary? No, because the two events are independent. Winning the coin flip did not alter the probability of either candidate drawing the winning slip of paper.

A survey by the American Automobile Association (AAA) revealed 60% of its members made airline reservations last year. Two members are selected at random. What is the probability both made airline reservations last year?

**SOLUTION**

The probability the first member made an airline reservation last year is .60, written  $P(R_1) = .60$ , where  $R_1$  refers to the fact that the first member made a reservation. The probability that the second member selected made a reservation is also .60, so  $P(R_2) = .60$ . Because the number of AAA members is very large, you may assume that  $R_1$  and  $R_2$  are independent. Consequently, using formula (5–5), the probability they both make a reservation is .36, found by:

$$P(R_1 \text{ and } R_2) = P(R_1)P(R_2) = (.60)(.60) = .36$$

All possible outcomes can be shown as follows.  $R$  means a reservation is made, and  $\sim R$  means no reservation is made.

With the probabilities and the complement rule, we can compute the joint probability of each outcome. For example, the probability that neither member makes a reservation is .16. Further, the probability of the first or the second member (special addition rule) making a reservation is .48 or (.24 + .24). You can also observe that the outcomes are mutually exclusive and collectively exhaustive. Therefore, the probabilities sum to 1.00.

| Outcomes            | Joint Probability  |
|---------------------|--------------------|
| $R_1 R_2$           | $(.60)(.60) = .36$ |
| $R_1 \sim R_2$      | $(.60)(.40) = .24$ |
| $\sim R_1 R_2$      | $(.40)(.60) = .24$ |
| $\sim R_1 \sim R_2$ | $(.40)(.40) = .16$ |
| Total               | 1.00               |

**SELF-REVIEW 5–5**



From experience, Teton Tire knows the probability is .95 that a particular XB-70 tire will last 60,000 miles before it becomes bald or fails. An adjustment is made on any tire that does not last 60,000 miles. You purchase four XB-70s. What is the probability all four tires will last at least 60,000 miles?

**General Rule of Multiplication**

If two events are not independent, they are referred to as **dependent**. To illustrate dependency, suppose there are 10 cans of soda in a cooler; 7 are regular and 3 are diet. A can is selected from the cooler. The probability of selecting a can of diet soda is 3/10, and the probability of selecting a can of regular soda is 7/10. Then a second can is selected from the cooler, without returning the first. The probability the second is diet depends on whether the first one selected was diet or not. The probability that the second is diet is:

- 2/9, if the first can is diet. (Only two cans of diet soda remain in the cooler.)
- 3/9, if the first can selected is regular. (All three diet sodas are still in the cooler.)

The fraction  $2/9$  (or  $3/9$ ) is called a **conditional probability** because its value is conditional on (dependent on) whether a diet or regular soda was the first selection from the cooler.

**CONDITIONAL PROBABILITY** The probability of a particular event occurring, given that another event has occurred.

In the general rule of multiplication, the conditional probability is required to compute the joint probability of two events that are not independent. For two events,  $A$  and  $B$ , that are not independent, the conditional probability is represented as  $P(B|A)$ , and expressed as the probability of  $B$  given  $A$ . Or the probability of  $B$  is conditional on the occurrence and effect of event  $A$ . Symbolically, the general rule of multiplication for two events that are not independent is:

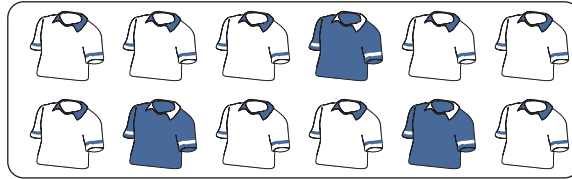
**GENERAL RULE OF MULTIPLICATION**

$$P(A \text{ and } B) = P(A)P(B|A)$$

**(5–6)**

► **EXAMPLE**

A golfer has 12 golf shirts in his closet. Suppose 9 of these shirts are white and the others blue. He gets dressed in the dark, so he just grabs a shirt and puts it on. He plays golf two days in a row and does not launder and return the used shirts to the closet. What is the likelihood both shirts selected are white?



**SOLUTION**

The event that the first shirt selected is white is  $W_1$ . The probability is  $P(W_1) = 9/12$  because 9 of the 12 shirts are white. The event that the second shirt selected is also white is identified as  $W_2$ . The conditional probability that the second shirt selected is white, given that the first shirt selected is also white, is  $P(W_2|W_1) = 8/11$ . Why is this so? Because after the first shirt is selected, there are only 11 shirts remaining in the closet and 8 of these are white. To determine the probability of 2 white shirts being selected, we use formula (5–6).

$$P(W_1 \text{ and } W_2) = P(W_1)P(W_2|W_1) = \left(\frac{9}{12}\right)\left(\frac{8}{11}\right) = .55$$

So the likelihood of selecting two shirts and finding them both to be white is .55.

We can extend the general rule of multiplication to more than two events. For three events  $A$ ,  $B$ , and  $C$ , the formula is:

$$P(A \text{ and } B \text{ and } C) = P(A)P(B|A)P(C|A \text{ and } B)$$

In the case of the golf shirt example, the probability of selecting three white shirts without replacement is:

$$P(W_1 \text{ and } W_2 \text{ and } W_3) = P(W_1)P(W_2|W_1)P(W_3|W_1 \text{ and } W_2) = \left(\frac{9}{12}\right)\left(\frac{8}{11}\right)\left(\frac{7}{10}\right) = .38$$

So the likelihood of selecting three shirts without replacement and all being white is .38.

## SELF-REVIEW 5-6



The board of directors of Tarbell Industries consists of eight men and four women. A four-member search committee is to be chosen at random to conduct a nationwide search for a new company president.

- What is the probability all four members of the search committee will be women?
- What is the probability all four members will be men?
- Does the sum of the probabilities for the events described in parts (a) and (b) equal 1? Explain.

### LO5-5

Compute probabilities using a contingency table.

## CONTINGENCY TABLES

Often we tally the results of a survey in a two-way table and use the results of this tally to determine various probabilities. We described this idea on page 106 in Chapter 4. To review, we refer to a two-way table as a **contingency table**.

**CONTINGENCY TABLE** A table used to classify sample observations according to two or more identifiable characteristics.

A contingency table is a cross-tabulation that simultaneously summarizes two variables of interest and their relationship. The level of measurement can be nominal. Below are several examples.

- One hundred fifty adults were asked their gender and the number of Facebook accounts they used. The following table summarizes the results.

| Facebook Accounts | Gender    |           | Total     |
|-------------------|-----------|-----------|-----------|
|                   | Men       | Women     |           |
| 0                 | 20        | 40        | 60        |
| 1                 | 40        | 30        | 70        |
| 2 or more         | <u>10</u> | <u>10</u> | <u>20</u> |
| Total             | 70        | 80        | 150       |

- The American Coffee Producers Association reports the following information on age and the amount of coffee consumed in a month.

| Age (Years) | Coffee Consumption |           |           | Total     |
|-------------|--------------------|-----------|-----------|-----------|
|             | Low                | Moderate  | High      |           |
| Under 30    | 36                 | 32        | 24        | 92        |
| 30 up to 40 | 18                 | 30        | 27        | 75        |
| 40 up to 50 | 10                 | 24        | 20        | 54        |
| 50 and over | <u>26</u>          | <u>24</u> | <u>29</u> | <u>79</u> |
| Total       | 90                 | 110       | 100       | 300       |

According to this table, each of the 300 respondents is classified according to two criteria: (1) age and (2) the amount of coffee consumed.

The following example shows how the rules of addition and multiplication are used when we employ contingency tables.



▶ **EXAMPLE**

Last month, the National Association of Theater Managers conducted a survey of 500 randomly selected adults. The survey asked respondents their age and the number of times they saw a movie in a theater. The results are summarized in Table 5–1.

**TABLE 5–1** Number of Movies Attended per Month by Age

| Movies per Month |       | Age                   |                      |                      | Total |
|------------------|-------|-----------------------|----------------------|----------------------|-------|
|                  |       | Less than 30<br>$B_1$ | 30 up to 60<br>$B_2$ | 60 or Older<br>$B_3$ |       |
| 0                | $A_1$ | 15                    | 50                   | 10                   | 75    |
| 1 or 2           | $A_2$ | 25                    | 100                  | 75                   | 200   |
| 3, 4, or 5       | $A_3$ | 55                    | 60                   | 60                   | 175   |
| 6 or more        | $A_4$ | 5                     | 15                   | 30                   | 50    |
| <b>Total</b>     |       | 100                   | 225                  | 175                  | 500   |

The association is interested in understanding the probabilities that an adult will see a movie in a theater, especially for adults 60 and older. This information is useful for making decisions regarding discounts on tickets and concessions for seniors.

Determine the probability of:

1. Selecting an adult who attended 6 or more movies per month.
2. Selecting an adult who attended 2 or fewer movies per month.
3. Selecting an adult who attended 6 or more movies per month **or** is 60 years of age or older.
4. Selecting an adult who attended 6 or more movies per month **given** the person is 60 years of age or older.
5. Selecting an adult who attended 6 or more movies per month **and** is 60 years of age or older.

Determine the independence of:

6. Number of movies per month attended and the age of the adult.

**SOLUTION**

Table 5–1 is called a contingency table. In a contingency table, an individual or an object is classified according to two criteria. In this example, a sampled adult is classified by age and by the number of movies attended per month. The rules of addition [formulas (5–2) and (5–4)] and the rules of multiplication [formulas (5–5) and (5–6)] allow us to answer the various probability questions based on the contingency table.

1. To find the probability that a randomly selected adult attended 6 or more movies per month, focus on the row labeled “6 or more” (also labeled  $A_4$ ) in Table 5–1. The table shows that 50 of the total of 500 adults are in this class. Using the empirical approach, the probability is computed:

$$P(6 \text{ or more}) = P(A_4) = \frac{50}{500} = .10$$

This probability indicates 10% of the 500 adults attend 6 or more movies per month.

2. To determine the probability of randomly selecting an adult who went to 2 or fewer movies per month, two outcomes must be combined: attending 0 movies per month and attending 1 or 2 movies per month. These two outcomes are mutually exclusive. That is, a person can only be classified as attending 0

## STATISTICS IN ACTION

A recent study by the National Collegiate Athletic Association (NCAA) reported that of 150,000 senior boys playing on their high school basketball team, 64 would make a professional team. To put it another way, the odds of a high school senior basketball player making a professional team are 1 in 2,344. From the same study:

1. The odds of a high school senior basketball player playing some college basketball are about 1 in 40.
2. The odds of a high school senior playing college basketball as a senior in college are about 1 in 60.
3. If you play basketball as a senior in college, the odds of making a professional team are about 1 in 37.5.

movies per month or 1 or 2 movies per month, not both. Because the two outcomes are mutually exclusive, we use the special rule of addition [formula (5-2)] by adding the probabilities of attending no movies and attending 1 or 2 movies:

$$P[(\text{attending 0}) \text{ or } (\text{attending 1 or 2})] = P(A_1) + P(A_2) = \left( \frac{75}{500} + \frac{200}{500} \right) = .55$$

So 55% of the adults in the sample attended 2 or fewer movies per month.

3. To determine the probability of randomly selecting an adult who went to “6 or more” movies per month or whose age is “60 or older,” we again use the rules of addition. However, in this case the outcomes are **not** mutually exclusive. Why is this? Because a person can attend 6 or more movies per month, be 60 or older, or be both. So the two groups are not mutually exclusive because it is possible that a person would be counted in both groups. To determine this probability, the general rule of addition [formula (5-4)] is used.

$$\begin{aligned} P[(6 \text{ or more}) \text{ or } (60 \text{ or older})] &= P(A_4) + P(B_3) - P(A_4 \text{ and } B_3) \\ &= \left( \frac{50}{500} + \frac{175}{500} - \frac{30}{500} \right) = .39 \end{aligned}$$

So 39% of the adults are either 60 or older, attend 6 or more movies per month, or both.

4. To determine the probability of selecting a person who attends 6 or more movies per month given that the person is 60 or older, focus only on the column labeled  $B_3$  in Table 5-1. That is, we are only interested in the 175 adults who are 60 or older. Of these 175 adults, 30 attended 6 or more movies. Using the general rule of multiplication [formula (5-6)]:

$$P[(6 \text{ or more}) \text{ given } (60 \text{ or older})] = P(A_4|B_3) = \frac{30}{175} = .17$$

Of the 500 adults, 17% of adults who are 60 or older attend 6 or more movies per month. This is called a conditional probability because the probability is based on the “condition” of being the age of 60 or older. Recall that in part (1), 10% of all adults attend 6 or more movies per month; here we see that 17% of adults who are 60 or older attend movies. This is valuable information for theater managers regarding the characteristics of their customers.

5. The probability a person attended 6 or more movies and is 60 or older is based on two conditions and they must both happen. That is, the two outcomes “6 or more movies” ( $A_4$ ) and “60 or older” ( $B_3$ ) must occur jointly. To find this joint probability, we use the special rule of multiplication [formula (5-6)].

$$P[(6 \text{ or more}) \text{ and } (60 \text{ or older})] = P(A_4 \text{ and } B_3) = P(A_4)P(B_3|A_4)$$

To compute the joint probability, first compute the simple probability of the first outcome,  $A_4$ , randomly selecting a person who attends 6 or more movies. To find the probability, refer to row  $A_4$  in Table 5-1. There are 50 of 500 adults that attended 6 or more movies. So  $P(A_4) = 50/500$ .

Next, compute the conditional probability  $P(B_3|A_4)$ . This is the probability of selecting an adult who is 60 or older given that the person attended 6 or more movies. The conditional probability is:

$$P[(60 \text{ or older}) \text{ given } (60 \text{ or more})] = P(B_3|A_4) = 30/50$$

Using these two probabilities, the joint probability that an adult attends 6 or more movies and is 60 or older is:

$$\begin{aligned} P[(6 \text{ or more}) \text{ and } (60 \text{ or older})] &= P(A_4 \text{ and } B_3) = P(A_4)P(B_3|A_4) \\ &= (50/500)(30/50) = .06 \end{aligned}$$

Based on the sample information from Table 5–1, the probability that an adult is both over 60 and attended 6 or more movies is 6%. It is important to know that the 6% is relative to all 500 adults.

Is there another way to determine this joint probability without using the special rule of multiplication formula? Yes. Look directly at the cell where row  $A_4$ , attends 6 or more movies, and column  $B_3$ , 60 or older, intersect. There are 30 adults in this cell that meet both criteria, so  $P(A_4 \text{ and } B_3) = 30/500 = .06$ . This is the same as computed with the formula.

6. Are the events independent? We can answer this question with the help of the results in part 4. In part 4 we found the probability of selecting an adult who was 60 or older given that the adult attended 6 or more movies was .17. If age is not a factor in movie attendance, then we would expect the probability of a person who is 30 or less that attended 6 or more movies to also be 17%. That is, the two conditional probabilities would be the same. The probability that an adult attends 6 or more movies per month given the adult is less than 30 years old is:

$$P[(6 \text{ or more}) \text{ given (less than 30)}] = \frac{5}{100} = .05$$

Because these two probabilities are not the same, the number of movies attended and age are not independent. To put it another way, for the 500 adults, age is related to the number of movies attended. In Chapter 15, we investigate this concept of independence in greater detail.

## SELF-REVIEW 5-7



Refer to Table 5–1 on page 136 to find the following probabilities.

- What is the probability of selecting an adult who is 30 up to 60 years old?
- What is the probability of selecting an adult who is under 60 years of age?
- What is the probability of selecting an adult who is less than 30 years old or attended no movies?
- What is the probability of selecting an adult who is less than 30 years old and went to no movies?

## Tree Diagrams

A **tree diagram** is a visual that is helpful in organizing and calculating probabilities for problems, similar to the previous example/solution. This type of problem involves several stages, and each stage is illustrated with a branch of the tree. The branches of a tree diagram are labeled with probabilities. We will use the information in Table 5–1 to show the construction of a tree diagram.

- We begin the construction by drawing a box with the variable age on the left to represent the root of the tree (see Chart 5–2).
- There are three main branches going out from the root. The upper branch represents the outcome that an adult is less than 30 years old. The branch is labeled with the probability,  $P(B_1) = 100/500$ . The next branch represents the outcome that adults are 30 up to 60 years old. This branch is labeled with the probability  $P(B_2) = 225/500$ . The remaining branch is labeled  $P(B_3) = 175/500$ .
- Four branches “grow” out of each of the four main branches. These branches represent the four categories of movies attended per month—0; 1 or 2; 3, 4, or 5; and 6 or more. The upper branches of the tree represent the conditional probabilities that an adult did not attend any movies given they are less than 30 years old. These are written  $P(A_1|B_1)$ ,  $P(A_2|B_1)$ ,  $P(A_3|B_1)$ , and  $P(A_4|B_1)$ , where  $A_1$  refers to attending no movies;  $A_2$  attending 1 or 2 movies per month;  $A_3$  attending 3, 4, or 5 movies

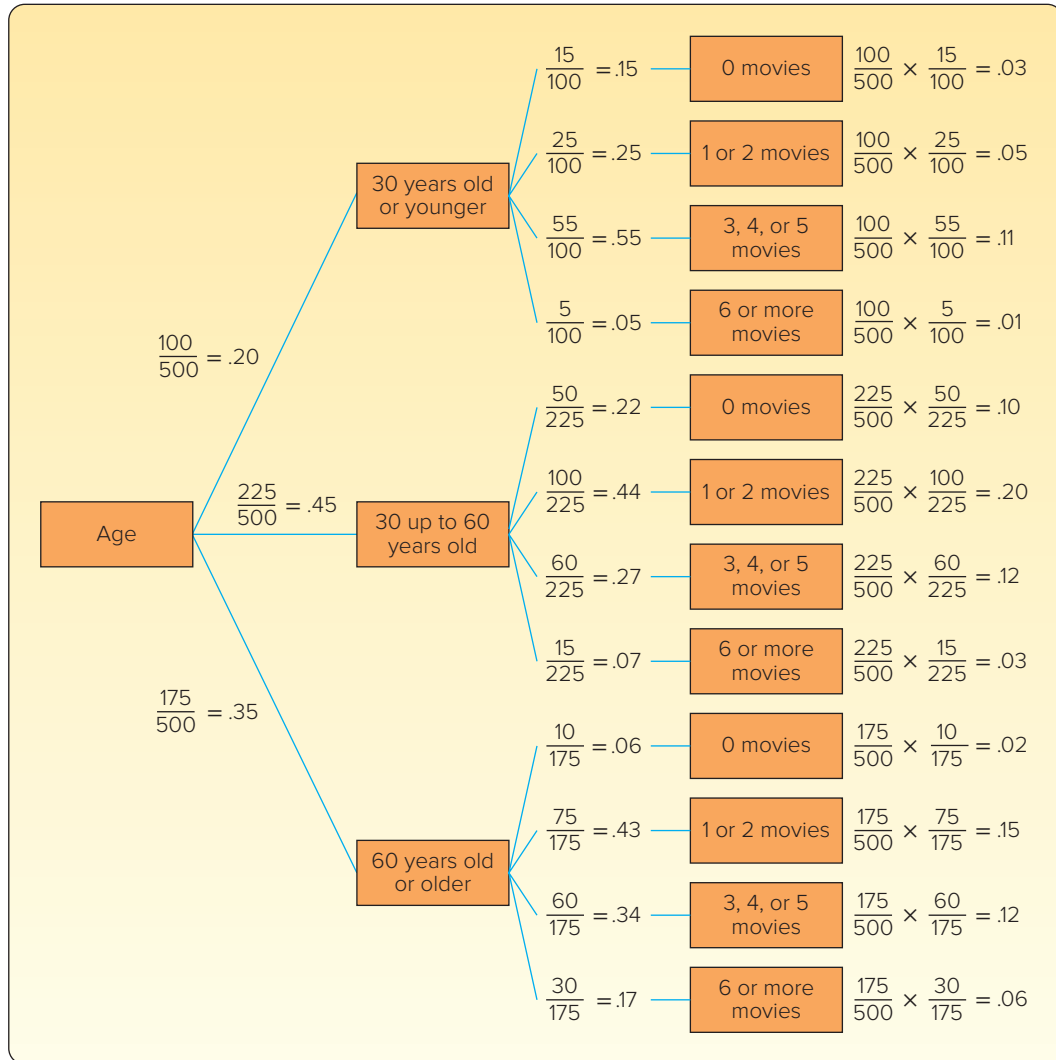


CHART 5-2 Tree Diagram Showing Age and Number of Movies Attended

per month; and  $A_4$  attending 6 or more movies per month. For the upper branch of the tree, these probabilities are  $15/100$ ,  $25/100$ ,  $55/100$ , and  $5/100$ . We write the conditional probabilities in a similar fashion on the other branches.

- Finally we determine the various joint probabilities. For the top branches, the events are an adult attends no movies per month and is 30 years old or younger; an adult attends 1 or 2 movies and is 30 years old or younger; an adult attends 3, 4, or 5 movies per month and is 30 years old or younger; and an adult attends 6 or more movies per month and is 30 years old or younger. These joint probabilities are shown on the right side of Chart 5-2. To explain, the joint probability that a randomly selected adult is less than 30 years old and attends 0 movies per month is:

$$P(B_1 \text{ and } A_1) = P(B_1)P(A_1|B_1) = \left(\frac{100}{500}\right)\left(\frac{15}{100}\right) = .03$$

The tree diagram summarizes all the probabilities based on the contingency table in Table 5-1. For example, the conditional probabilities show that the 60-and-older group has the highest percentage, 17%, attending 6 or movies per month. The

30-to-60-year-old group has the highest percentage, 22%, of seeing no movies per month. Based on the joint probabilities, 20% of the adults sampled attend 1 or 2 movies per month and are 30 up to 60 years of age. As you can see, there are many observations that we can make based on the information presented in the tree diagram.

## SELF-REVIEW 5-8



Consumers were surveyed on the relative number of visits to a Sears store (often, occasional, and never) and if the store was located in an enclosed mall (yes and no). When variables are measured nominally, such as these data, the results are usually summarized in a contingency table.

| Visits     | Enclosed Mall |     | Total |
|------------|---------------|-----|-------|
|            | Yes           | No  |       |
| Often      | 60            | 20  | 80    |
| Occasional | 25            | 35  | 60    |
| Never      | 5             | 50  | 55    |
|            | 90            | 105 | 195   |

What is the probability of selecting a shopper who:

- Visited a Sears store often?
- Visited a Sears store in an enclosed mall?
- Visited a Sears store in an enclosed mall or visited a Sears store often?
- Visited a Sears store often, given that the shopper went to a Sears store in an enclosed mall?

In addition:

- Are the number of visits and the enclosed mall variables independent?
- What is the probability of selecting a shopper who visited a Sears store often and it was in an enclosed mall?
- Draw a tree diagram and determine the various joint probabilities.

## EXERCISES

- Suppose  $P(A) = .40$  and  $P(B|A) = .30$ . What is the joint probability of  $A$  and  $B$ ?
- Suppose  $P(X_1) = .75$  and  $P(Y_2|X_1) = .40$ . What is the joint probability of  $X_1$  and  $Y_2$ ?
- A local bank reports that 80% of its customers maintain a checking account, 60% have a savings account, and 50% have both. If a customer is chosen at random, what is the probability the customer has either a checking or a savings account? What is the probability the customer does not have either a checking or a savings account?
- All Seasons Plumbing has two service trucks that frequently need repair. If the probability the first truck is available is .75, the probability the second truck is available is .50, and the probability that both trucks are available is .30, what is the probability neither truck is available?
- FILE** Refer to the following table.

| Second Event | First Event |       |       | Total |
|--------------|-------------|-------|-------|-------|
|              | $A_1$       | $A_2$ | $A_3$ |       |
| $B_1$        | 2           | 1     | 3     | 6     |
| $B_2$        | 1           | 2     | 1     | 4     |
| Total        | 3           | 3     | 4     | 10    |

- a. Determine  $P(A_1)$ .  
 b. Determine  $P(B_1|A_2)$ .  
 c. Determine  $P(B_2 \text{ and } A_3)$ .
28. Three defective electric toothbrushes were accidentally shipped to a drugstore by Cleanbrush Products along with 17 nondefective ones.
- What is the probability the first two electric toothbrushes sold will be returned to the drugstore because they are defective?
  - What is the probability the first two electric toothbrushes sold will not be defective?
29. **FILE** Each salesperson at Puchett, Sheets, and Hogan Insurance Agency is rated either below average, average, or above average with respect to sales ability. Each salesperson also is rated with respect to his or her potential for advancement—either fair, good, or excellent. These traits for the 500 salespeople were cross-classified into the following table.

| Sales Ability | Potential for Advancement |      |           |
|---------------|---------------------------|------|-----------|
|               | Fair                      | Good | Excellent |
| Below average | 16                        | 12   | 22        |
| Average       | 45                        | 60   | 45        |
| Above average | 93                        | 72   | 135       |

- What is this table called?
  - What is the probability a salesperson selected at random will have above average sales ability and excellent potential for advancement?
  - Construct a tree diagram showing all the probabilities, conditional probabilities, and joint probabilities.
30. An investor owns three common stocks. Each stock, independent of the others, has equally likely chances of (1) increasing in value, (2) decreasing in value, or (3) remaining the same value. List the possible outcomes of this experiment. Estimate the probability at least two of the stocks increase in value.
31. **FILE** A survey of 545 college students asked: What is your favorite winter sport? And, what type of college do you attend? The results are summarized below.

| College Type      | Favorite Winter Sport |        |             | Total |
|-------------------|-----------------------|--------|-------------|-------|
|                   | Snowboarding          | Skiing | Ice Skating |       |
| Junior college    | 68                    | 41     | 46          | 155   |
| Four-year college | 84                    | 56     | 70          | 210   |
| Graduate school   | 59                    | 74     | 47          | 180   |
| Total             | 211                   | 171    | 163         | 545   |

Using these 545 students as the sample, a student from this study is randomly selected.

- What is the probability of selecting a student whose favorite sport is skiing?
  - What is the probability of selecting a junior-college student?
  - If the student selected is a four-year-college student, what is the probability that the student prefers ice skating?
  - If the student selected prefers snowboarding, what is the probability that the student is in junior college?
  - If a graduate student is selected, what is the probability that the student prefers skiing or ice skating?
32. If you ask three strangers about their birthdays, what is the probability (a) All were born on Wednesday? (b) All were born on different days of the week? (c) None was born on Saturday?

**LO5-6**

Determine the number of outcomes using principles of counting.

## PRINCIPLES OF COUNTING

If the number of possible outcomes in an experiment is small, it is relatively easy to count them. There are six possible outcomes, for example, resulting from the roll of a die, namely:



If, however, there are a large number of possible outcomes, such as the number of heads and tails for an experiment with 10 tosses, it would be tedious to count all the possibilities. They could have all heads, one head and nine tails, two heads and eight tails, and so on. To facilitate counting, we describe three formulas: the multiplication formula (not to be confused with the multiplication *rule* described earlier in the chapter), the permutation formula, and the combination formula.

### The Multiplication Formula

We begin with the **multiplication formula**.

**MULTIPLICATION FORMULA** If there are  $m$  ways of doing one thing and  $n$  ways of doing another thing, there are  $m \times n$  ways of doing both.

In terms of a formula:

**MULTIPLICATION FORMULA** Total number of arrangements =  $(m)(n)$  (5-7)

This can be extended to more than two events. For three events  $m$ ,  $n$ , and  $o$ :

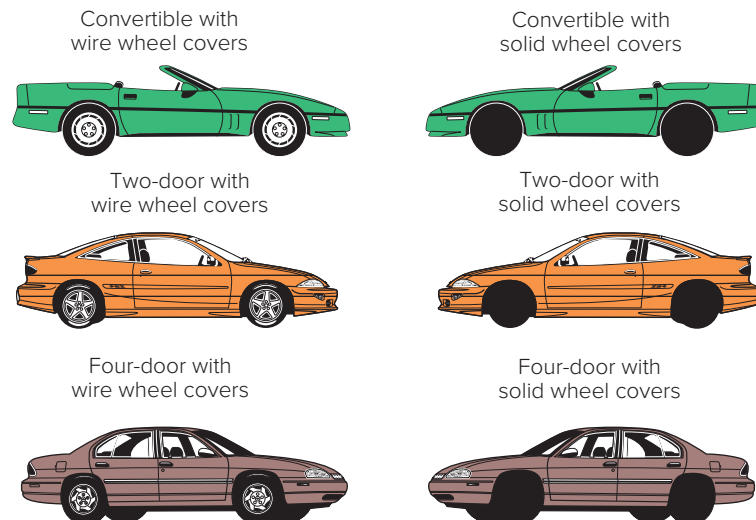
$$\text{Total number of arrangements} = (m)(n)(o)$$

### EXAMPLE

An automobile dealer wants to advertise that for \$29,999 you can buy a convertible, a two-door sedan, or a four-door model with your choice of either wire wheel covers or solid wheel covers. Based on the number of models and wheel covers, how many different vehicles can the dealer offer?

### SOLUTION

Of course, the dealer could determine the total number of different cars by picturing and counting them. There are six.



We can employ the multiplication formula as a check (where  $m$  is the number of models and  $n$  the wheel cover type). From formula (5–7):

$$\text{Total possible arrangements} = (m)(n) = (3)(2) = 6$$

It was not difficult to count all the possible model and wheel cover combinations in this example. Suppose, however, that the dealer decided to offer eight models and six types of wheel covers. It would be tedious to picture and count all the possible alternatives. Instead, the multiplication formula can be used. In this case, there are  $(m)(n) = (8)(6) = 48$  possible arrangements.

Note in the preceding applications of the multiplication formula that there were *two or more groupings from which you made selections*. The automobile dealer, for example, offered a choice of models and a choice of wheel covers. If a home builder offered you four different exterior styles of a home to choose from and three interior floor plans, the multiplication formula would be used to find how many different arrangements were possible. There are 12 possibilities.

## SELF-REVIEW 5–9



1. The Women's Shopping Network on cable TV offers sweaters and slacks for women. The sweaters and slacks are offered in coordinating colors. If sweaters are available in five colors and the slacks are available in four colors, how many different outfits can be advertised?
2. Pioneer manufactures three models of Wi-Fi Internet radios, two MP3 docking stations, four different sets of speakers, and three CD carousel changers. When the four types of components are sold together, they form a "system." How many different systems can the electronics firm offer?

## The Permutation Formula

The multiplication formula is applied to find the number of possible arrangements for two or more groups. In contrast, we use the **permutation formula** to find the number of possible arrangements when there is a single group of objects. Illustrations of this type of problem are:

- Three electronic parts—a transistor, an LED, and a synthesizer—are assembled into a plug-in component for an HDTV. The parts can be assembled in any order. How many different ways can the three parts be assembled?
- A machine operator must make four safety checks before starting his machine. It does not matter in which order the checks are made. In how many different ways can the operator make the checks?

One order for the first illustration might be the transistor first, the LED second, and the synthesizer third. This arrangement is called a **permutation**.

**PERMUTATION** Any arrangement of  $r$  objects selected from a single group of  $n$  possible objects.

Note that the arrangements  $a b c$  and  $b a c$  are different permutations. The formula to count the total number of different permutations is:

**PERMUTATION FORMULA**

$${}_n P_r = \frac{n!}{(n-r)!}$$

(5–8)



where:

$n$  is the total number of objects.

$r$  is the number of objects selected.

Before we solve the two problems illustrated, the permutations and combinations (to be discussed shortly) use a notation called *n factorial*. It is written  $n!$  and means the product of  $n(n-1)(n-2)(n-3)\cdots(1)$ . For instance,  $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$ .

Many of your calculators have a button with  $x!$  that will perform this calculation for you. It will save you a great deal of time. For example the Texas Instruments Pro Scientific calculator has the following key:



It is the “third function,” so check your users’ manual or the Internet for instructions.

The factorial notation can also be canceled when the same number appears in both the numerator and the denominator, as shown below.

$$\frac{6!3!}{4!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1(3 \cdot 2 \cdot 1)}{4 \cdot 3 \cdot 2 \cdot 1} = 180$$

By definition, zero factorial, written  $0!$ , is 1. That is,  $0! = 1$ .

### EXAMPLE

Referring to the group of three electronic parts that are to be assembled in any order, in how many different ways can they be assembled?

### SOLUTION

There are three electronic parts to be assembled, so  $n = 3$ . Because all three are to be inserted into the plug-in component,  $r = 3$ . Solving using formula (5–8) gives:

$${}_n P_r = \frac{n!}{(n-r)!} = \frac{3!}{(3-3)!} = \frac{3!}{0!} = \frac{3!}{1} = 6$$

We can check the number of permutations arrived at by using the permutation formula. We determine how many “spaces” have to be filled and the possibilities for each “space.” In the problem involving three electronic parts, there are three locations in the plug-in unit for the three parts. There are three possibilities for the first place, two for the second (one has been used up), and one for the third, as follows:

$$(3)(2)(1) = 6 \text{ permutations}$$

The six ways in which the three electronic parts, lettered  $A, B, C$ , can be arranged are:

$ABC \quad BAC \quad CAB \quad ACB \quad BCA \quad CBA$

In the previous example, we selected and arranged all the objects, that is  $n = r$ . In many cases, only some objects are selected and arranged from the  $n$  possible objects. We explain the details of this application in the following example.

### EXAMPLE

The Fast Media Company is producing a one-minute video advertisement. In the production process, eight different video segments were made. To make the one-minute ad, they can only select three of the eight segments. How many different ways can the eight video segments be arranged in the three spaces available in the ad?

**SOLUTION**

There are eight possibilities for the first available space in the ad, seven for the second space (one has been used up), and six for the third space. Thus:

$$(8)(7)(6) = 336$$

That is, there are a total of 336 different possible arrangements. This could also be found by using formula (5–8). If  $n = 8$  video segments and  $r = 3$  spaces available, the formula leads to

$${}_n P_r = \frac{n!}{(n-r)!} = \frac{8!}{(8-3)!} = \frac{8!}{5!} = \frac{(8)(7)(6)\cancel{5!}}{\cancel{5!}} = 336$$

**The Combination Formula**

If the order of the selected objects is *not* important, any selection is called a **combination**. Logically, the number of combinations is always less than the number of permutations. The formula to count the number of  $r$  object combinations from a set of  $n$  objects is:

**COMBINATION FORMULA**

$${}_n C_r = \frac{n!}{r!(n-r)!}$$

**(5–9)**

For example, if executives Able, Baker, and Chauncy are to be chosen as a committee to negotiate a merger, there is only one possible combination of these three; the committee of Able, Baker, and Chauncy is the same as the committee of Baker, Chauncy, and Able. Using the combination formula:

$${}_n C_r = \frac{n!}{r!(n-r)!} = \frac{3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1(1)} = 1$$

**EXAMPLE**

The Grand 16 movie theater uses teams of three employees to work the concession stand each evening. There are seven employees available to work each evening. How many different teams can be scheduled to staff the concession stand?

**SOLUTION**

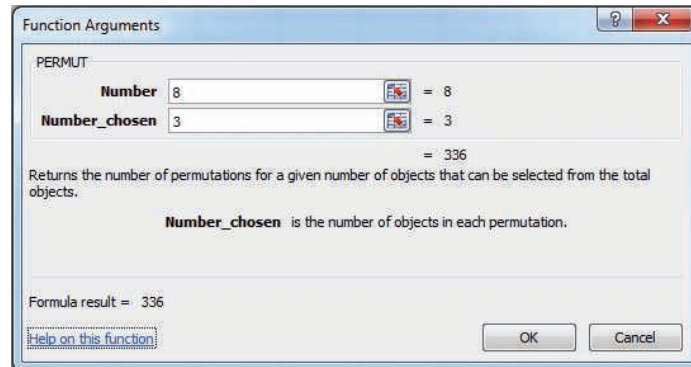
According to formula (5–9), there are 35 combinations, found by

$${}_7 C_3 = \frac{n!}{r!(n-r)!} = \frac{7!}{3!(7-3)!} = \frac{7!}{3!4!} = 35$$

The seven employees taken three at a time would create the possibility of 35 different teams.

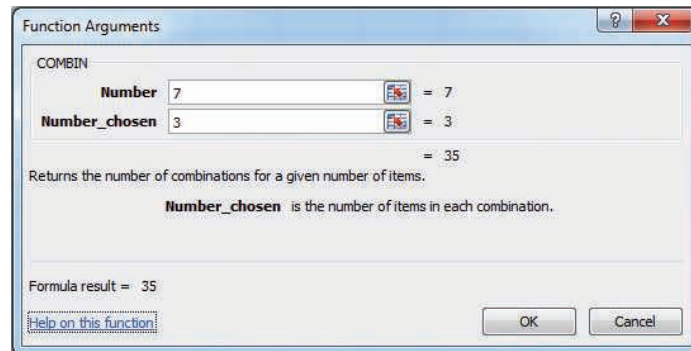
When the number of permutations or combinations is large, the calculations are tedious. Computer software and handheld calculators have “functions” to compute these numbers. The results using the PERMUT function in Excel applied to the selection of three

video segments for the eight available at the Fast Media Company is shown below. There are a total of 336 arrangements.



Source: Microsoft Excel

Below is the result using the COMBIN function in Excel applied to the number of possible teams of three selected from seven employees at the Grand 16 movie theater. There are 35 possible teams of three.



Source: Microsoft Excel

## SELF-REVIEW 5-10



- A musician wants to write a score based on only five chords: B-flat, C, D, E, and G. However, only three chords out of the five will be used in succession, such as C, B-flat, and E. Repetitions, such as B-flat, B-flat, and E, will not be permitted.
  - How many permutations of the five chords, taken three at a time, are possible?
  - Using formula (5-8), how many permutations are possible?
- The 10 numbers 0 through 9 are to be used in code groups of four to identify an item of clothing. Code 1083 might identify a blue blouse, size medium; the code group 2031 might identify a pair of pants, size 18; and so on. Repetitions of numbers are not permitted. That is, the same number cannot be used twice (or more) in a total sequence. For example, 2256, 2562, or 5559 would not be permitted. How many different code groups can be designed?
- In the preceding example/solution involving the Grand 16 movie theater, there were 35 possible teams of three taken from seven employees.
  - Use formula (5-9) to show this is true.
  - The manager of the theater wants to plan for staffing the concession stand with teams of five employees on the weekends to serve the larger crowds. From the seven employees, how many teams of five employees are possible?
- In a lottery game, three numbers are randomly selected from a tumbler of balls numbered 1 through 50.
  - How many permutations are possible?
  - How many combinations are possible?

## EXERCISES

33. Solve the following:
  - a.  $40!/35!$
  - b.  ${}_7P_4$
  - c.  ${}_5C_2$
34. Solve the following:
  - a.  $20!/17!$
  - b.  ${}_9P_3$
  - c.  ${}_7C_2$
35. A pollster randomly selected 4 of 10 available people. How many different groups of 4 are possible?
36. A telephone number consists of seven digits, the first three representing the exchange. How many different telephone numbers are possible within the 537 exchange?
37. An overnight express company must include five cities on its route. How many different routes are possible, assuming that it does not matter in which order the cities are included in the routing?
38. A representative of the Environmental Protection Agency (EPA) wants to select samples from 10 landfills. The director has 15 landfills from which she can collect samples. How many different samples are possible?
39. Sam Snead's restaurant in Conway, South Carolina, offers an early bird special from 4–6 p.m. each weekday evening. If each patron selects a Starter Selection (4 options), an Entrée (8 options), and a Dessert (3 options), how many different meals are possible?
40. A company is creating three new divisions, and seven managers are eligible to be appointed head of a division. How many different ways could the three new heads be appointed? Hint: Assume the division assignment makes a difference.

## CHAPTER SUMMARY

- I. A probability is a value between 0 and 1 inclusive that represents the likelihood a particular event will happen.
  - A. An experiment is the observation of some activity or the act of taking some measurement.
  - B. An outcome is a particular result of an experiment.
  - C. An event is the collection of one or more outcomes of an experiment.
- II. There are three definitions of probability.
  - A. The classical definition applies when there are  $n$  equally likely outcomes to an experiment.
  - B. The empirical definition occurs when the number of times an event happens is divided by the number of observations.
  - C. A subjective probability is based on whatever information is available.
- III. Two events are mutually exclusive if by virtue of one event happening the other cannot happen.
- IV. Events are independent if the occurrence of one event does not affect the occurrence of another event.
- V. The rules of addition refer to the probability that any of two or more events can occur.
  - A. The special rule of addition is used when events are mutually exclusive.
 
$$P(A \text{ or } B) = P(A) + P(B) \quad (5-2)$$
  - B. The general rule of addition is used when the events are not mutually exclusive.
 
$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \quad (5-4)$$
  - C. The complement rule is used to determine the probability of an event happening by subtracting the probability of the event not happening from 1.
 
$$P(A) = 1 - P(\sim A) \quad (5-3)$$

- VI.** The rules of multiplication are applied when two or more events occur simultaneously.  
**A.** The special rule of multiplication refers to events that are independent.

$$P(A \text{ and } B) = P(A)P(B) \quad (5-5)$$

- B.** The general rule of multiplication refers to events that are not independent.

$$P(A \text{ and } B) = P(A)P(B|A) \quad (5-6)$$

- C.** A joint probability is the likelihood that two or more events will happen at the same time.  
**D.** A conditional probability is the likelihood that an event will happen, given that another event has already happened.

- VII.** There are three counting rules that are useful in determining the number of outcomes in an experiment.

- A.** The multiplication rule states that if there are  $m$  ways one event can happen and  $n$  ways another event can happen, then there are  $mn$  ways the two events can happen.

$$\text{Number of arrangements} = (m)(n) \quad (5-7)$$

- B.** A permutation is an arrangement in which the order of the objects selected from a specific pool of objects is important.

$${}_n P_r = \frac{n!}{(n-r)!} \quad (5-8)$$

- C.** A combination is an arrangement where the order of the objects selected from a specific pool of objects is not important.

$${}_n C_r = \frac{n!}{r!(n-r)!} \quad (5-9)$$

## PRONUNCIATION KEY

| SYMBOL                | MEANING   | PRONUNCIATION        |
|-----------------------|---|----------------------|
| $P(A)$                | Probability of $A$                              | $P$ of $A$           |
| $P(\sim A)$           | Probability of not $A$                          | $P$ of not $A$       |
| $P(A \text{ and } B)$ | Probability of $A$ and $B$                      | $P$ of $A$ and $B$   |
| $P(A \text{ or } B)$  | Probability of $A$ or $B$                       | $P$ of $A$ or $B$    |
| $P(A B)$              | Probability of $A$ given $B$ has happened       | $P$ of $A$ given $B$ |
| ${}_n P_r$            | Permutation of $n$ items selected $r$ at a time | $Pnr$                |
| ${}_n C_r$            | Combination of $n$ items selected $r$ at a time | $Cnr$                |

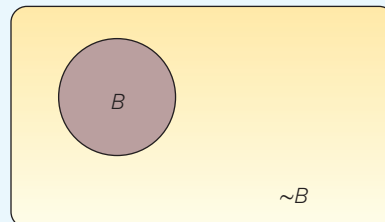
## CHAPTER EXERCISES

- 41.** The marketing research department at PepsiCo plans to survey teenagers about a newly developed soft drink. Each will be asked to compare it with his or her favorite soft drink.
- What is the experiment?
  - What is one possible event?
- 42.** The number of times a particular event occurred in the past is divided by the number of occurrences. What is this approach to probability called?
- 43.** The probability that the cause and the cure for all cancers will be discovered before the year 2020 is .20. What viewpoint of probability does this statement illustrate?
- 44.** **FILE** Berdine's Chicken Factory has several stores in the Hilton Head, South Carolina, area. When interviewing applicants for server positions, the owner would like to include information on the amount of tip a server can expect to earn per check (or bill).

A study of 500 recent checks indicated the server earned the following amounts in tips per 8-hour shift.

| Amount of Tip   | Number |
|-----------------|--------|
| \$0 up to \$ 20 | 200    |
| 20 up to 50     | 100    |
| 50 up to 100    | 75     |
| 100 up to 200   | 75     |
| 200 or more     | 50     |
| Total           | 500    |

- a. What is the probability of a tip of \$200 or more?
  - b. Are the categories “\$0 up to \$20,” “\$20 up to \$50,” and so on considered mutually exclusive?
  - c. If the probabilities associated with each outcome were totaled, what would that total be?
  - d. What is the probability of a tip of up to \$50?
  - e. What is the probability of a tip of less than \$200?
45. Winning all three “Triple Crown” races is considered the greatest feat of a pedigree racehorse. After a successful Kentucky Derby, Corn on the Cob is a heavy favorite at 2 to 1 odds to win the Preakness Stakes.
    - a. If he is a 2 to 1 favorite to win the Belmont Stakes as well, what is his probability of winning the Triple Crown?
    - b. What do his chances for the Preakness Stakes have to be in order for him to be “even money” to earn the Triple Crown?
  46. The first card selected from a standard 52-card deck is a king.
    - a. If it is returned to the deck, what is the probability that a king will be drawn on the second selection?
    - b. If the king is not replaced, what is the probability that a king will be drawn on the second selection?
    - c. What is the probability that a king will be selected on the first draw from the deck and another king on the second draw (assuming that the first king was not replaced)?
  47. Armco, a manufacturer of traffic light systems, found 95% of the newly developed systems lasted 3 years before failing to change signals properly.
    - a. If a city purchased four of these systems, what is the probability all four systems would operate properly for at least 3 years?
    - b. Which rule of probability does this illustrate?
    - c. Using letters to represent the four systems, write an equation to show how you arrived at the answer to part (a).
  48. Refer to the following picture.



- a. What is the picture called?
- b. What rule of probability is illustrated?
- c.  $B$  represents the event of choosing a family that receives welfare payments. What does  $P(B) + P(\sim B)$  equal?

- 49.** In a management trainee program at Claremont Enterprises, 80% of the trainees are female and 20% male. Ninety percent of the females attended college, and 78% of the males attended college.
- A management trainee is selected at random. What is the probability that the person selected is a female who did not attend college?
  - Are gender and attending college independent? Why?
  - Construct a tree diagram showing all the probabilities, conditional probabilities, and joint probabilities.
  - Do the joint probabilities total 1.00? Why?
- 50.** Assume the likelihood that any flight on Delta Airlines arrives within 15 minutes of the scheduled time is .90. We randomly selected a Delta flight on four different days.
- What is the likelihood all four of the selected flights arrived within 15 minutes of the scheduled time?
  - What is the likelihood that none of the selected flights arrived within 15 minutes of the scheduled time?
  - What is the likelihood at least one of the selected flights did not arrive within 15 minutes of the scheduled time?
- 51.** There are 100 employees at Kiddie Carts International. Fifty-seven of the employees are hourly workers, 40 are supervisors, 2 are secretaries, and the remaining employee is the president. Suppose an employee is selected:
- What is the probability the selected employee is an hourly worker?
  - What is the probability the selected employee is either an hourly worker or a supervisor?
  - Refer to part (b). Are these events mutually exclusive?
  - What is the probability the selected employee is neither an hourly worker nor a supervisor?
- 52.** DJ LeMahieu of the Colorado Rockies had the highest batting average in the 2016 Major League Baseball season. His average was .348. So assume the probability of getting a hit is .348 for each time he batted. In a particular game, assume he batted three times.
- This is an example of what type of probability?
  - What is the probability of getting three hits in a particular game?
  - What is the probability of not getting any hits in a game?
  - What is the probability of getting at least one hit?
- 53.** Four women's college basketball teams are participating in a single-elimination holiday basketball tournament. If one team is favored in its semifinal match by odds of 2 to 1 and another squad is favored in its contest by odds of 3 to 1, what is the probability that:
- Both favored teams win their games?
  - Neither favored team wins its game?
  - At least one of the favored teams wins its game?
- 54.** There are three clues labeled "daily double" on the game show *Jeopardy*. If three equally matched contenders play, what is the probability that:
- A single contestant finds all three "daily doubles"?
  - The returning champion gets all three of the "daily doubles"?
  - Each of the players selects precisely one of the "daily doubles"?
- 55.** Brooks Insurance Inc. wishes to offer life insurance to men age 60 via the Internet. Mortality tables indicate the likelihood of a 60-year-old man surviving another year is .98. If the policy is offered to five men age 60:
- What is the probability all five men survive the year?
  - What is the probability at least one does not survive?
- 56.** Forty percent of the homes constructed in the Quail Creek area include a security system. Three homes are selected at random:
- What is the probability all three of the selected homes have a security system?
  - What is the probability none of the three selected homes has a security system?
  - What is the probability at least one of the selected homes has a security system?
  - Did you assume the events to be dependent or independent?
- 57.** Refer to Exercise 56, but assume there are 10 homes in the Quail Creek area and 4 of them have a security system. Three homes are selected at random:
- What is the probability all three of the selected homes have a security system?
  - What is the probability none of the three selected homes has a security system?

- c. What is the probability at least one of the selected homes has a security system?  
 d. Did you assume the events to be dependent or independent?
58. There are 20 families living in the Willbrook Farms Development. Of these families, 10 prepared their own federal income taxes for last year, 7 had their taxes prepared by a local professional, and the remaining 3 were done by H&R Block.
- a. What is the probability of selecting a family that prepared their own taxes?  
 b. What is the probability of selecting two families, both of which prepared their own taxes?  
 c. What is the probability of selecting three families, all of which prepared their own taxes?  
 d. What is the probability of selecting two families, neither of which had their taxes prepared by H&R Block?
59. The board of directors of Saner Automatic Door Company consists of 12 members, 3 of whom are women. A new policy and procedures manual is to be written for the company. A committee of three is randomly selected from the board to do the writing.
- a. What is the probability that all members of the committee are men?  
 b. What is the probability that at least one member of the committee is a woman?
60. **FILE** A recent survey reported in *Bloomberg Businessweek* dealt with the salaries of CEOs at large corporations and whether company shareholders made money or lost money.

|                         | CEO Paid More<br>Than \$1 Million | CEO Paid Less<br>Than \$1 Million | Total |
|-------------------------|-----------------------------------|-----------------------------------|-------|
| Shareholders made money | 2                                 | 11                                | 13    |
| Shareholders lost money | 4                                 | 3                                 | 7     |
| Total                   | 6                                 | 14                                | 20    |

- If a company is randomly selected from the list of 20 studied, what is the probability:
- a. The CEO made more than \$1 million?  
 b. The CEO made more than \$1 million or the shareholders lost money?  
 c. The CEO made more than \$1 million given the shareholders lost money?  
 d. Of selecting two CEOs and finding they both made more than \$1 million?
61. Althoff and Roll, an investment firm in Augusta, Georgia, advertises extensively in the *Augusta Morning Gazette*, the newspaper serving the region. The *Gazette* marketing staff estimates that 60% of Althoff and Roll's potential market read the newspaper. It is further estimated that 85% of those who read the *Gazette* remember the Althoff and Roll advertisement.
- a. What percent of the investment firm's potential market sees and remembers the advertisement?  
 b. What percent of the investment firm's potential market sees, but does not remember, the advertisement?
62. An Internet company located in Southern California has season tickets to the Los Angeles Lakers basketball games. The company president always invites one of the four vice presidents to attend games with him, and claims he selects the person to attend at random. One of the four vice presidents has not been invited to attend any of the last five Lakers home games. What is the likelihood this could be due to chance?
63. A computer-supply retailer purchased a batch of 1,000 CD-R disks and attempted to format them for a particular application. There were 857 perfect CDs, 112 CDs were usable but had bad sectors, and the remainder could not be used at all.
- a. What is the probability a randomly chosen CD is not perfect?  
 b. If the disk is not perfect, what is the probability it cannot be used at all?
64. An investor purchased 100 shares of Fifth Third Bank stock and 100 shares of Santee Electric Cooperative stock. The probability the bank stock will appreciate over a year is .70. The probability the electric utility will increase over the same period is .60. Assume the two events are independent.
- a. What is the probability both stocks appreciate during the period?  
 b. What is the probability the bank stock appreciates but the utility does not?  
 c. What is the probability at least one of the stocks appreciates?



- 65.** With each purchase of a large pizza at Tony's Pizza, the customer receives a coupon that can be scratched to see if a prize will be awarded. The probability of winning a free soft drink is 0.10, and the probability of winning a free large pizza is 0.02. You plan to eat lunch tomorrow at Tony's. What is the probability:
- That you will win either a large pizza or a soft drink?
  - That you will not win a prize?
  - That you will not win a prize on three consecutive visits to Tony's?
  - That you will win at least one prize on one of your next three visits to Tony's?
- 66.** For the daily lottery game in Illinois, participants select three numbers between 0 and 9. A number cannot be selected more than once, so a winning ticket could be, say, 307 but not 337. Purchasing one ticket allows you to select one set of numbers. The winning numbers are announced on TV each night.
- How many different outcomes (three-digit numbers) are possible?
  - If you purchase a ticket for the game tonight, what is the likelihood you will win?
  - Suppose you purchase three tickets for tonight's drawing and select a different number for each ticket. What is the probability that you will not win with any of the tickets?
- 67.** Several years ago, Wendy's advertised that there are 256 different ways to order your hamburger. You may choose to have, or omit, any combination of the following on your hamburger: mustard, ketchup, onion, pickle, tomato, relish, mayonnaise, and lettuce. Is the advertisement correct? Show how you arrive at your answer.
- 68.** Recent surveys indicate 60% of tourists to China visited the Forbidden City, the Temple of Heaven, the Great Wall, and other historical sites in or near Beijing. Forty percent visited Xi'an with its magnificent terra-cotta soldiers, horses, and chariots, which lay buried for over 2,000 years. Thirty percent of the tourists went to both Beijing and Xi'an. What is the probability that a tourist visited at least one of these places?
- 69.** A new chewing gum has been developed that is helpful to those who want to stop smoking. If 60% of those people chewing the gum are successful in stopping smoking, what is the probability that in a group of four smokers using the gum at least one quits smoking?
- 70.** Reynolds Construction Company has agreed not to erect all "look-alike" homes in a new subdivision. Five exterior designs are offered to potential home buyers. The builder has standardized three interior plans that can be incorporated in any of the five exteriors. How many different ways can the exterior and interior plans be offered to potential home buyers?
- 71.** A new sports car model has defective brakes 15% of the time and a defective steering mechanism 5% of the time. Let's assume (and hope) that these problems occur independently. If one or the other of these problems is present, the car is called a "lemon." If both of these problems are present, the car is a "hazard." Your instructor purchased one of these cars yesterday. What is the probability it is:
- A lemon?
  - A hazard?
- 72.** The state of Maryland has license plates with three numbers followed by three letters. How many different license plates are possible?
- 73.** There are four people being considered for the position of chief executive officer of Dalton Enterprises. Three of the applicants are over 60 years of age. Two are female, of which only one is over 60.
- What is the probability that a candidate is over 60 and female?
  - Given that the candidate is male, what is the probability he is younger than 60?
  - Given that the person is over 60, what is the probability the person is female?
- 74.** Tim Bleckie is the owner of Bleckie Investment and Real Estate Company. The company recently purchased four tracts of land in Holly Farms Estates and six tracts in Newburg Woods. The tracts are all equally desirable and sell for about the same amount.
- What is the probability that the next two tracts sold will be in Newburg Woods?
  - What is the probability that of the next four sold, at least one will be in Holly Farms?
  - Are these events independent or dependent?
- 75.** A computer password consists of four characters. The characters can be one of the 26 letters of the alphabet. Each character may be used more than once. How many different passwords are possible?

- 76.** A case of 24 cans contains 1 can that is contaminated. Three cans are to be chosen randomly for testing.
- How many different combinations of three cans could be selected?
  - What is the probability that the contaminated can is selected for testing?
- 77.** A puzzle in the newspaper presents a matching problem. The names of 10 U.S. presidents are listed in one column, and their vice presidents are listed in random order in the second column. The puzzle asks the reader to match each president with his vice president. If you make the matches randomly, how many matches are possible? What is the probability all 10 of your matches are correct?
- 78.** Two components,  $A$  and  $B$ , operate in series. Being in series means that for the system to operate, both components  $A$  and  $B$  must work. Assume the two components are independent. What is the probability the system works under these conditions? The probability  $A$  works is .90 and the probability  $B$  functions is also .90.
- 79.** You take a trip by air that involves three independent flights. If there is an 80% chance each specific leg of the trip is on time, what is the probability all three flights arrive on time?
- 80.** The probability a D-Link network server is down is .05. If you have three independent servers, what is the probability that at least one of them is operational?
- 81.** Twenty-two percent of all light emitting diode (LED) displays are manufactured by Samsung. What is the probability that in a collection of three independent LED HDTV purchases, at least one is a Samsung?

## DATA ANALYTICS

(The data for these exercises are available in Connect.)

- 82. FILE** Refer to the North Valley Real Estate data, which report information on homes sold during the last year.
- Sort the data into a table that shows the number of homes that have a pool versus the number that don't have a pool in each of the five townships. If a home is selected at random, compute the following probabilities.
    - The home has a pool.
    - The home is in Township 1 or has a pool.
    - Given that it is in Township 3, that it has a pool.
    - The home has a pool and is in Township 3.
  - Sort the data into a table that shows the number of homes that have a garage attached versus those that don't in each of the five townships. If a home is selected at random, compute the following probabilities:
    - The home has a garage attached.
    - The home does not have a garage attached, given that it is in Township 5.
    - The home has a garage attached and is in Township 3.
    - The home does not have a garage attached or is in Township 2.
- 83. FILE** Refer to the Baseball 2016 data, which reports information on the 30 Major League Baseball teams for the 2016 season. Set up three variables:
- Divide the teams into two groups, those that had a winning season and those that did not. That is, create a variable to count the teams that won 81 games or more, and those that won 80 or less.
  - Create a new variable for attendance, using three categories: attendance less than 2.0 million, attendance of 2.0 million up to 3.0 million, and attendance of 3.0 million or more.
  - Create a variable that shows the teams that play in a stadium less than 20 years old versus one that is 20 years old or more.
- Answer the following questions.
- Create a table that shows the number of teams with a winning season versus those with a losing season by the three categories of attendance. If a team is selected at random, compute the following probabilities:
    - The team had a winning season.
    - The team had a winning season or attendance of more than 3.0 million.
    - The team had a winning season given attendance was more than 3.0 million.
    - The team had a winning season and attracted fewer than 2.0 million fans.

- b. Create a table that shows the number of teams with a winning season versus those that play in new or old stadiums. If a team is selected at random, compute the following probabilities:
    1. Selecting a team with a stadium that is at least 20 years old.
    2. The likelihood of selecting a team with a winning record and playing in a new stadium.
    3. The team had a winning record or played in a new stadium.
84. **FILE** Refer to the Lincolnville school bus data. Set up a variable that divides the age of the buses into three groups: new (less than 5 years old), medium (at least 5 but less than 10 years), and old (10 or more years). The median maintenance cost is \$4,179. Based on this value, create a variable for those less than or equal to the median (low maintenance) and those more than the median (high maintenance cost). Finally, develop a table to show the relationship between maintenance cost and age of the bus.
- a. What percentage of the buses are less than five years old?
  - b. What percentage of the buses less than five years old have low maintenance costs?
  - c. What percentage of the buses ten or more years old have high maintenance costs?
  - d. Does maintenance cost seem to be related to the age of the bus? Hint: Compare the maintenance cost of the old buses with the cost of the new buses. Would you conclude maintenance cost is independent of the age?

## PRACTICE TEST

### Part 1—Objective

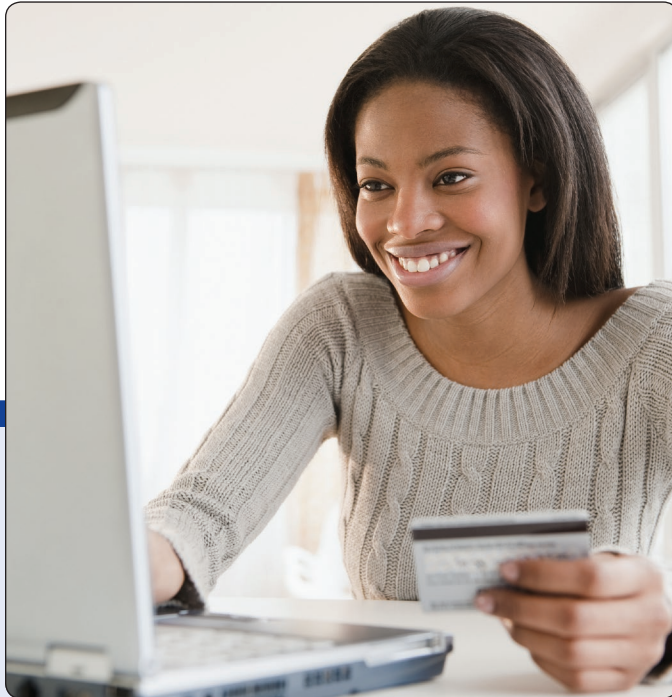
1. A \_\_\_\_\_ is a value between zero and one, inclusive, describing the relative chance or likelihood an event will occur.
2. An \_\_\_\_\_ is a process that leads to the occurrence of one and only one of several possible outcomes.
3. An \_\_\_\_\_ is a collection of one or more outcomes of an experiment.
4. Using the \_\_\_\_\_ viewpoint, the probability of an event happening is the fraction of the time similar events happened in the past.
5. Using the \_\_\_\_\_ viewpoint, an individual evaluates the available opinions and information and then estimates or assigns the probability.
6. Using the \_\_\_\_\_ viewpoint, the probability of an event happening is computed by dividing the number of favorable outcomes by the number of possible outcomes.
7. If several events are described as \_\_\_\_\_, then the occurrence of one event means that none of the other events can occur at the same time.
8. If an experiment has a set of events that includes every possible outcome, then the set of events is described as \_\_\_\_\_.
9. If two events  $A$  and  $B$  are \_\_\_\_\_, the special rule of addition states that the probability of one or the other events occurring equals the sum of their probabilities.
10. The \_\_\_\_\_ is used to determine the probability of an event occurring by subtracting the probability of the event not occurring from 1.
11. A probability that measures the likelihood two or more events will happen concurrently is called a \_\_\_\_\_.
12. The special rule of multiplication requires that two events  $A$  and  $B$  are \_\_\_\_\_.

### Part 2—Problems

1. Fred Friendly, CPA, has a stack of 20 tax returns to complete before the April 15th deadline. Of the 20 tax returns, 12 are from individuals, 5 are from businesses, and 3 are from charitable organizations. He randomly selects two returns. What is the probability that:
  - a. Both are businesses?
  - b. At least one is a business?
2. Fred exercises regularly. His fitness log for the last 12 months shows that he jogged 30% of the days, rode his bike 20% of the days, and did both on 12% of the days. What is the probability that Fred would do at least one of these two types of exercises on any given day?
3. Fred works in a tax office with four other CPAs. There are five parking spots beside the office. If they all drive to work, how many different ways can the cars belonging to the CPAs be arranged in the five spots?

# Discrete Probability Distributions

# 6



©JGI/Jamie Grill/Getty Images RF

- ▲ **RECENT STATISTICS SUGGEST** that 15% of those who visit a retail site on the Web make a purchase. A retailer wished to verify this claim. To do so, she selected a sample of 16 “hits” to her site and found that 4 had actually made a purchase. What is the likelihood of exactly four purchases? How many purchases should she expect? What is the likelihood that four or more “hits” result in a purchase? (See Exercise 43 and **LO6-4**.)

## LEARNING OBJECTIVES

---

*When you have completed this chapter, you will be able to:*

- LO6-1** Identify the characteristics of a probability distribution.
- LO6-2** Distinguish between discrete and continuous random variables.
- LO6-3** Compute the mean, variance, and standard deviation of a discrete probability distribution.
- LO6-4** Explain the assumptions of the binomial distribution and apply it to calculate probabilities.
- LO6-5** Explain the assumptions of the Poisson distribution and apply it to calculate probabilities.

## INTRODUCTION

Chapters 2 through 4 are devoted to descriptive statistics. We describe raw data by organizing the data into a frequency distribution and portraying the distribution in tables, graphs, and charts. Also, we compute a measure of location—such as the arithmetic mean, median, or mode—to locate a typical value near the center of the distribution. The range and the standard deviation are used to describe the spread in the data. These chapters focus on describing *something that has already happened*.

Starting with Chapter 5, the emphasis changes—we begin examining *something that could happen*. We note that this facet of statistics is called *statistical inference*. The objective is to make inferences (statements) about a population based on a number of observations, called a sample, selected from the population. In Chapter 5, we state that a probability is a value between 0 and 1 inclusive, and we examine how probabilities can be combined using rules of addition and multiplication.

This chapter begins the study of **probability distributions**. A probability distribution is like a relative frequency distribution. However, instead of describing the past, it is used to provide estimates of the likelihood of future events. Probability distributions can be described by measures of location and dispersion, so we show how to compute a distribution's mean, variance, and standard deviation. We also discuss two frequently occurring discrete probability distributions: the binomial and Poisson.

### LO6-1

Identify the characteristics of a probability distribution.

## WHAT IS A PROBABILITY DISTRIBUTION?

A probability distribution defines or describes the likelihoods for a range of possible future outcomes. For example, Spalding Golf Products Inc. assembles golf clubs with three components: a club head, a shaft, and a grip. From experience, 5% of the shafts received from their Asian supplier are defective. As part of Spalding's statistical process control, they inspect 20 shafts from each arriving shipment. From experience, we know that the probability of a defective shaft is 5%. Therefore, in a sample of 20 shafts, we would expect 1 shaft to be defective and the other 19 shafts to be acceptable. But, by using a probability distribution, we can completely describe the range of possible outcomes. For example, we would know the probability that none of the 20 shafts are defective, or that 2, or 3, or 4, or continuing up to 20 shafts in the sample are defective. Given the small probability of a defective shaft, the probability distribution would show that there is a very small probability of 4 or more defective shafts.

**PROBABILITY DISTRIBUTION** A listing of all the outcomes of an experiment and the probability associated with each outcome.

The important characteristics of a probability distribution are:

### CHARACTERISTICS OF A PROBABILITY DISTRIBUTION

1. The probability of a particular outcome is between 0 and 1 inclusive.
2. The outcomes are mutually exclusive.
3. The list of outcomes is exhaustive. So the sum of the probabilities of the outcomes is equal to 1.

How can we generate a probability distribution? The following example will explain.

**EXAMPLE**

Suppose we are interested in the number of heads showing face up on three tosses of a coin. This is the experiment. The possible results are zero heads, one head, two heads, and three heads. What is the probability distribution for the number of heads?

**SOLUTION**

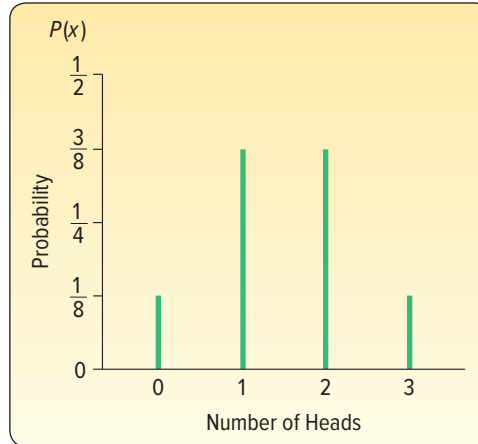
There are eight possible outcomes. A tail might appear face up on the first toss, another tail on the second toss, and another tail on the third toss of the coin. Or we might get a tail, tail, and head, in that order. We use the multiplication formula for counting outcomes (5–7). There are  $(2)(2)(2)$  or 8 possible results. These results are shown in the following table.

| Possible Result | Coin Toss |        |       | Number of Heads |
|-----------------|-----------|--------|-------|-----------------|
|                 | First     | Second | Third |                 |
| 1               | T         | T      | T     | 0               |
| 2               | T         | T      | H     | 1               |
| 3               | T         | H      | T     | 1               |
| 4               | T         | H      | H     | 2               |
| 5               | H         | T      | T     | 1               |
| 6               | H         | T      | H     | 2               |
| 7               | H         | H      | T     | 2               |
| 8               | H         | H      | H     | 3               |

Note that the outcome “zero heads” occurred only once, “one head” occurred three times, “two heads” occurred three times, and the outcome “three heads” occurred only once. That is, “zero heads” happened one out of eight times. Thus, the probability of zero heads is one-eighth, the probability of one head is three-eighths, and so on. The probability distribution is shown in Table 6–1. Because one of these outcomes must happen, the total of the probabilities of all possible events is 1.000. This is always true. The same information is shown in Chart 6–1.

**TABLE 6–1** Probability Distribution for the Events of Zero, One, Two, and Three Heads Showing Face Up on Three Tosses of a Coin

| Number of Heads,<br>$x$ | Probability of Outcome,<br>$P(x)$ |
|-------------------------|-----------------------------------|
| 0                       | $\frac{1}{8} = .125$              |
| 1                       | $\frac{3}{8} = .375$              |
| 2                       | $\frac{3}{8} = .375$              |
| 3                       | $\frac{1}{8} = .125$              |
| Total                   | $\frac{8}{8} = 1.000$             |



**CHART 6-1** Graphical Presentation of the Number of Heads Resulting from Three Tosses of a Coin and the Corresponding Probability

Refer to the coin-tossing example in Table 6-1. We write the probability of  $x$  as  $P(x)$ . So the probability of zero heads is  $P(0 \text{ heads}) = .125$ , and the probability of one head is  $P(1 \text{ head}) = .375$ , and so forth. The sum of these mutually exclusive probabilities is 1; that is, from Table 6-1,  $.125 + .375 + .375 + .125 = 1.00$ .

## SELF-REVIEW 6-1



The possible outcomes of an experiment involving the roll of a six-sided die are a one-spot, a two-spot, a three-spot, a four-spot, a five-spot, and a six-spot.

- Develop a probability distribution for the number of possible spots.
- Portray the probability distribution graphically.
- What is the sum of the probabilities?

### LO6-2

Distinguish between discrete and continuous random variables.

## RANDOM VARIABLES

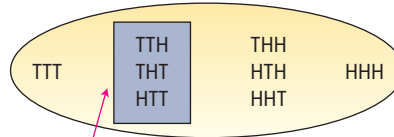
In any experiment of chance, the outcomes occur randomly. So it is often called a *random variable*. For example, rolling a single die is an experiment: Any one of six possible outcomes can occur. Some experiments result in outcomes that are measured with quantitative variables (such as dollars, weight, or number of children), and other experimental outcomes are measured with qualitative variables (such as color or religious preference). A few examples will further illustrate what is meant by a **random variable**.

- The number of employees absent from the day shift on Monday might be 0, 1, 2, 3, . . . The number absent is the random variable.
- The hourly wage of a sample of 50 plumbers in Jacksonville, Florida. The hourly wage is the random variable.
- The number of defective lightbulbs produced in an hour at the Cleveland Electric Company Inc.
- The grade level (Freshman, Sophomore, Junior, or Senior) of the members of the St. James High School Varsity girls' basketball team. The grade level is the random variable; notice that it is a qualitative variable.
- The number of participants in the 2017 New York City Marathon.
- The daily number of drivers charged with driving under the influence of alcohol in Brazoria County, Texas, last month.

A random variable is defined as follows:

**RANDOM VARIABLE** A variable measured or observed as the result of an experiment. By chance, the variable can have different values.

In Chapter 5, we defined the terms *experiment*, *outcome*, and *event*. Consider the example we just described regarding the experiment of tossing a fair coin three times. In this case the *random variable* is the number of heads that appear in the three tosses. There are eight possible outcomes to this experiment. These outcomes are shown in the following diagram.

Possible *outcomes* for three coin tosses

The *event* {one head} occurs and the *random variable*  $x = 1$ .

So, one possible outcome is that a tail appears on each toss: TTT. This single outcome would describe the event of zero heads appearing in three tosses. Another possible outcome is a head followed by two tails: HTT. If we wish to determine the event of exactly one head appearing in the three tosses, we must consider the three possible outcomes: TTH, THT, and HTT. These three outcomes describe the event of exactly one head appearing in three tosses.

In this experiment, the random variable is the number of heads in three tosses. The random variable can have four different values, 0, 1, 2, or 3. The outcomes of the experiment are unknown. But, using probability, we can compute the probability of a single head in three tosses as  $3/8$  or 0.375. As shown in Chapter 5, the probability of each value of the random variable can be computed to create a probability distribution for the random variable, number of heads in three tosses of a coin.

There are two types of random variables: *discrete* or *continuous*.

## Discrete Random Variable

A discrete random variable can assume only a certain number of separated values. For example, the Bank of the Carolinas counts the number of credit cards carried for a group of customers. The data are summarized with the following relative frequency table.

| Number of Credit Cards | Relative Frequency |
|------------------------|--------------------|
| 0                      | .03                |
| 1                      | .10                |
| 2                      | .18                |
| 3                      | .21                |
| 4 or more              | <u>.48</u>         |
| Total                  | 1.00               |

In this frequency table, the number of cards carried is the **discrete random variable**.

**DISCRETE RANDOM VARIABLE** A random variable that can assume only certain clearly separated values.

A discrete random variable can, in some cases, assume fractional or decimal values. To be a discrete random variable, these values must be separated—that is, have distance between them. As an example, a department store offers coupons with discounts of 10%, 15%, and 25%. In terms of probability, we could compute the probability that a customer would use a 10% coupon versus a 15% or 25% coupon.



## Continuous Random Variable

On the other hand, a **continuous random variable** can assume an infinite number of values within a given range. It is measured on a continuous interval or ratio scale. Examples include:

- The times of commercial flights between Atlanta and Los Angeles are 4.67 hours, 5.13 hours, and so on. The random variable is the time in hours and is measured on a continuous scale of time.
- The annual snowfall in Minneapolis, Minnesota. The random variable is the amount of snow, measured on a continuous scale.

**CONTINUOUS RANDOM VARIABLE** A random variable that may assume an infinite number of values within a given range.

As with discrete random variables, the likelihood of a continuous random variable can be summarized with a **probability distribution**. For example, with a probability distribution for the flight time between Atlanta and Los Angeles, we could say that there is a probability of 0.90 that the flight will be less than 4.5 hours. This also implies that there is a probability of 0.10 that the flight will be more than 4.5 hours. With a probability of snowfall in Minneapolis, we could say that there is probability of 0.25 that the annual snowfall will exceed 48 inches. This also implies that there is a probability of 0.75 that annual snowfall will be less than 48 inches. Notice that these examples refer to a continuous range of values.

### LO6-3

Compute the mean, variance, and standard deviation of a discrete probability distribution.

## THE MEAN, VARIANCE, AND STANDARD DEVIATION OF A DISCRETE PROBABILITY DISTRIBUTION

In Chapter 3, we discussed measures of location and variation for a frequency distribution. The mean reports the central location of the data, and the variance describes the spread in the data. In a similar fashion, a probability distribution is summarized by its mean and variance. We identify the mean of a probability distribution by the lowercase Greek letter mu ( $\mu$ ) and the standard deviation by the lowercase Greek letter sigma ( $\sigma$ ).

### Mean

The mean is a typical value used to represent the central location of a probability distribution. It also is the long-run average value of the random variable. The mean of a probability distribution is also referred to as its **expected value**. It is a weighted average where the possible values of a random variable are weighted by their corresponding probabilities of occurrence.

The mean of a discrete probability distribution is computed by the formula:

**MEAN OF A PROBABILITY DISTRIBUTION**  $\mu = \sum[xP(x)]$  **(6-1)**

where  $P(x)$  is the probability of a particular value  $x$ . In other words, multiply each  $x$  value by its probability of occurrence, and then add these products.

### Variance and Standard Deviation

The mean is a typical value used to summarize a discrete probability distribution. However, it does not describe the amount of spread (variation) in a distribution. The variance does this. The formula for the variance of a probability distribution is:

**VARIANCE OF A PROBABILITY DISTRIBUTION**  $\sigma^2 = \sum[(x - \mu)^2P(x)]$  **(6-2)**

The computational steps are:

1. Subtract the mean from each value of the random variable, and square this difference.
2. Multiply each squared difference by its probability.
3. Sum the resulting products to arrive at the variance.

The standard deviation,  $\sigma$ , is found by taking the positive square root of  $\sigma^2$ ; that is,  $\sigma = \sqrt{\sigma^2}$ .

An example will help explain the details of the calculation and interpretation of the mean and standard deviation of a probability distribution.

**EXAMPLE**



©Don Mason/Getty Images RF

John Ragsdale sells new cars for Pelican Ford. John usually sells the largest number of cars on Saturday. He has developed the following probability distribution for the number of cars he expects to sell on a particular Saturday.

| Number of Cars Sold, $x$ | Probability, $P(x)$ |
|--------------------------|---------------------|
| 0                        | .1                  |
| 1                        | .2                  |
| 2                        | .3                  |
| 3                        | .3                  |
| 4                        | .1                  |
|                          | <u>1.0</u>          |

1. What type of distribution is this?
2. On a typical Saturday, how many cars does John expect to sell?
3. What is the variance of the distribution?

**SOLUTION**

1. This is a discrete probability distribution for the random variable called “number of cars sold.” Note that John expects to sell only within a certain range of cars; he does not expect to sell 5 cars or 50 cars. Further, he cannot sell half a car. He can sell only 0, 1, 2, 3, or 4 cars. Also, the outcomes are mutually exclusive—he cannot sell a total of both 3 and 4 cars on the same Saturday. The sum of the possible outcomes total 1. Hence, these circumstances qualify as a probability distribution.
2. The mean number of cars sold is computed by weighting the number of cars sold by the probability of selling that number and adding or summing the products, using formula (6–1):

$$\begin{aligned} \mu &= \sum[xP(x)] \\ &= 0(.1) + 1(.2) + 2(.3) + 3(.3) + 4(.1) \\ &= 2.1 \end{aligned}$$

These calculations are summarized in the following table.

| Number of Cars Sold, $x$ | Probability, $P(x)$ | $x \cdot P(x)$                |
|--------------------------|---------------------|-------------------------------|
| 0                        | .1                  | 0.0                           |
| 1                        | .2                  | 0.2                           |
| 2                        | .3                  | 0.6                           |
| 3                        | .3                  | 0.9                           |
| 4                        | .1                  | .4                            |
|                          | <u>1.0</u>          | <u><math>\mu = 2.1</math></u> |

How do we interpret a mean of 2.1? This value indicates that, over a large number of Saturdays, John Ragsdale expects to sell a mean of 2.1 cars a day. Of course, it is not possible for him to sell *exactly* 2.1 cars on any particular Saturday. However, the expected value can be used to predict the arithmetic mean number of cars sold on Saturdays in the long run. For example, if John works 50 Saturdays during a year, he can expect to sell (50)(2.1) or 105 cars, just on Saturdays. Thus, the mean is sometimes called the expected value.

3. The following table illustrates the steps to calculate the variance using formula (6–2). The first two columns repeat the probability distribution. In column three, the mean is subtracted from each value of the random variable. In column four, the differences from column three are squared. In the fifth column, each squared difference in column four is multiplied by the corresponding probability. The variance is the sum of the values in column five.

| Number of Cars Sold,<br>$x$ | Probability,<br>$P(x)$ | $(x - \mu)$ | $(x - \mu)^2$ | $(x - \mu)^2 P(x)$ |
|-----------------------------|------------------------|-------------|---------------|--------------------|
| 0                           | .1                     | $0 - 2.1$   | 4.41          | 0.441              |
| 1                           | .2                     | $1 - 2.1$   | 1.21          | 0.242              |
| 2                           | .3                     | $2 - 2.1$   | 0.01          | 0.003              |
| 3                           | .3                     | $3 - 2.1$   | 0.81          | 0.243              |
| 4                           | .1                     | $4 - 2.1$   | 3.61          | 0.361              |
|                             |                        |             |               | $\sigma^2 = 1.290$ |

Recall that the standard deviation,  $\sigma$ , is the positive square root of the variance. In this example,  $\sqrt{\sigma^2} = \sqrt{1.290} = 1.136$  cars. How do we apply a standard deviation of 1.136 cars? If salesperson Rita Kirsch also sold a mean of 2.1 cars on Saturdays, and the standard deviation in her sales was 1.91 cars, we would conclude that there is more variability in the Saturday sales of Ms. Kirsch than in those of Mr. Ragsdale (because  $1.91 > 1.136$ ).

## SELF-REVIEW 6-2



The Pizza Palace offers three sizes of cola. The smallest size sells for \$1.99, the medium for \$2.49, and the large for \$2.89. Thirty percent of the drinks sold are small, 50% are medium, and 20% are large. Create a probability distribution for the random variable price and answer the following questions.

- Is this a discrete probability distribution? Indicate why or why not.
- Compute the mean amount charged for a cola.
- What is the variance in the amount charged for a cola? The standard deviation?

## EXERCISES

1. **FILE** Compute the mean and variance of the following discrete probability distribution.

| $x$ | $P(x)$ |
|-----|--------|
| 0   | .2     |
| 1   | .4     |
| 2   | .3     |
| 3   | .1     |

2. **FILE** Compute the mean and variance of the following discrete probability distribution.

| $x$ | $P(x)$ |
|-----|--------|
| 2   | .5     |
| 8   | .3     |
| 10  | .2     |

3. **FILE** Compute the mean and variance of the following probability distribution.

| $x$ | $P(x)$ |
|-----|--------|
| 5   | .1     |
| 10  | .3     |
| 15  | .2     |
| 20  | .4     |

4. Which of these variables are discrete and which are continuous random variables?
- The number of new accounts established by a salesperson in a year.
  - The time between customer arrivals to a bank ATM.
  - The number of customers in Big Nick's barber shop.
  - The amount of fuel in your car's gas tank.
  - The number of minorities on a jury.
  - The outside temperature today.
5. **FILE** The information below is the number of daily emergency service calls made by the volunteer ambulance service of Walterboro, South Carolina, for the last 50 days. To explain, there were 22 days on which there were two emergency calls, and 9 days on which there were three emergency calls.

| Number of Calls | Frequency |
|-----------------|-----------|
| 0               | 8         |
| 1               | 10        |
| 2               | 22        |
| 3               | 9         |
| 4               | 1         |
| Total           | 50        |

- Convert this information on the number of calls to a probability distribution.
  - Is this an example of a discrete or continuous probability distribution?
  - What is the mean number of emergency calls per day?
  - What is the standard deviation of the number of calls made daily?
6. **FILE** The director of admissions at Kinzua University in Nova Scotia estimated the distribution of student admissions for the fall semester on the basis of past experience. What is the expected number of admissions for the fall semester? Compute the variance and the standard deviation of the number of admissions.

| Admissions | Probability |
|------------|-------------|
| 1,000      | .6          |
| 1,200      | .3          |
| 1,500      | .1          |

7. **FILE** Levinson's Department Store is having a special sale this weekend. Customers charging purchases of more than \$50 to their store credit card will be given a special Levinson's Lottery card. The customer will scratch off the card, which will indicate the amount to be taken off the total amount of the purchase. Listed below are the amount of the prize and the percent of the time that amount will be deducted from the total amount of the purchase.

| Prize Amount | Probability |
|--------------|-------------|
| \$ 10        | .50         |
| 25           | .40         |
| 50           | .08         |
| 100          | .02         |

- a. What is the mean amount deducted from the total purchase amount?
  - b. What is the standard deviation of the amount deducted from the total purchase?
8. **FILE** The Downtown Parking Authority of Tampa, Florida, reported the following information for a sample of 250 customers on the number of hours cars are parked and the amount they are charged.

| Number of Hours | Frequency | Amount Charged |
|-----------------|-----------|----------------|
| 1               | 20        | \$ 3           |
| 2               | 38        | 6              |
| 3               | 53        | 9              |
| 4               | 45        | 12             |
| 5               | 40        | 14             |
| 6               | 13        | 16             |
| 7               | 5         | 18             |
| 8               | 36        | 20             |
|                 | 250       |                |

- a. Convert the information on the number of hours parked to a probability distribution. Is this a discrete or a continuous probability distribution?
- b. Find the mean and the standard deviation of the number of hours parked. How long is a typical customer parked?
- c. Find the mean and the standard deviation of the amount charged.

**LO6-4**

Explain the assumptions of the binomial distribution and apply it to calculate probabilities.

## BINOMIAL PROBABILITY DISTRIBUTION

The **binomial probability distribution** is a widely occurring discrete probability distribution. To describe experimental outcomes with a binomial distribution, there are four requirements. The first requirement is there are only two possible outcomes on a particular experimental trial. For example, on a test, a true/false question is either answered correctly or incorrectly. In a resort, a housekeeping supervisor reviews an employee's work and evaluates it as acceptable or unacceptable. A key characteristic of the two outcomes is that they must be mutually exclusive. This means that the answer to a true/false question must be either correct or incorrect but cannot be both correct and incorrect at the same time. Another example is the outcome of a sales call. Either a customer purchases or does not purchase the product, but the sale cannot result in both outcomes. Frequently, we refer to the two possible outcomes of a binomial experiment as a "success" and a "failure." However, this distinction does not imply that one outcome is good and the other is bad, only that there are two mutually exclusive outcomes.

The second binomial requirement is that the random variable is the number of successes for a fixed and known number of trials. For example, we flip a coin five times and count the number of times a head appears in the five flips, we randomly select 10 employees and count the number who are older than 50 years of age, or we randomly select 20 boxes of Kellogg's Raisin Bran and count the number that weigh more than the amount indicated on the package. In each example, we count the number of successes from the fixed number of trials.

A third requirement is that we know the probability of a success and it is the same for each trial. Three examples are:

- For a test with 10 true/false questions, we know there are 10 trials and the probability of correctly guessing the answer for any of the 10 trials is 0.5. Or, for a test with 20 multiple-choice questions with four options and only one correct answer, we know that there are 20 trials and the probability of randomly guessing the correct answer for each of the 20 trials is 0.25.

- Bones Albaugh is a Division I college basketball player who makes 70% of his foul shots. If he has five opportunities in tonight's game, the likelihood he will be successful on each of the five attempts is 0.70.
- In a recent poll, 18% of adults indicated a Snickers bar was their favorite candy bar. We select a sample of 15 adults and ask each for his or her favorite candy bar. The likelihood a Snickers bar is the answer for each adult is 0.18.



©David Madison/Digital Vision/Getty Images RF

The final requirement of a binomial probability distribution is that each trial is *independent* of any other trial. Independent means there is no pattern to the trials. The outcome of a particular trial does not affect the outcome of any other trial. Two examples are:

- A young family has two children, both boys. The probability of a third birth being a boy is still .50. That is, the gender of the third child is independent of the gender of the other two.
- Suppose 20% of the patients served in the emergency room at Waccamaw Hospital do not have insurance. If the second patient served on the afternoon shift today did not have insurance, that does not affect the probability the third, the tenth, or any of the other patients will or will not have insurance.

#### BINOMIAL PROBABILITY EXPERIMENT

1. An outcome on each trial of an experiment is classified into one of two mutually exclusive categories—a success or a failure.
2. The random variable is the number of successes in a fixed number of trials.
3. The probability of success is the same for each trial.
4. The trials are independent, meaning that the outcome of one trial does not affect the outcome of any other trial.

## How is a Binomial Probability Computed?

To construct a particular binomial probability, we use (1) the number of trials and (2) the probability of success on each trial. For example, if the Hannah Landscaping Company plants 10 Norfolk pine trees today knowing that 90% of these trees survive, we can compute the binomial probability that exactly 8 trees survive. In this case, the number of trials is the 10 trees, the probability of success is .90, and the number of successes is eight. In fact, we can compute a binomial probability for any number of successes from 0 to 10 surviving trees.

A binomial probability is computed by the formula:

#### BINOMIAL PROBABILITY FORMULA

$$P(x) = {}_n C_x \pi^x (1 - \pi)^{n-x} \quad (6-3)$$

where:

$C$  denotes a combination.

$n$  is the number of trials.

$x$  is the random variable defined as the number of successes.

$\pi$  is the probability of a success on each trial.

We use the Greek letter  $\pi$  (pi) to denote a binomial population parameter. Do not confuse it with the mathematical constant 3.1416.

► **EXAMPLE**

There are five flights daily from Pittsburgh via American Airlines into the Bradford Regional Airport in Bradford, Pennsylvania. Suppose the probability that any flight arrives late is .20. What is the probability that none of the flights are late today? What is the probability that exactly one of the flights is late today?

**SOLUTION**

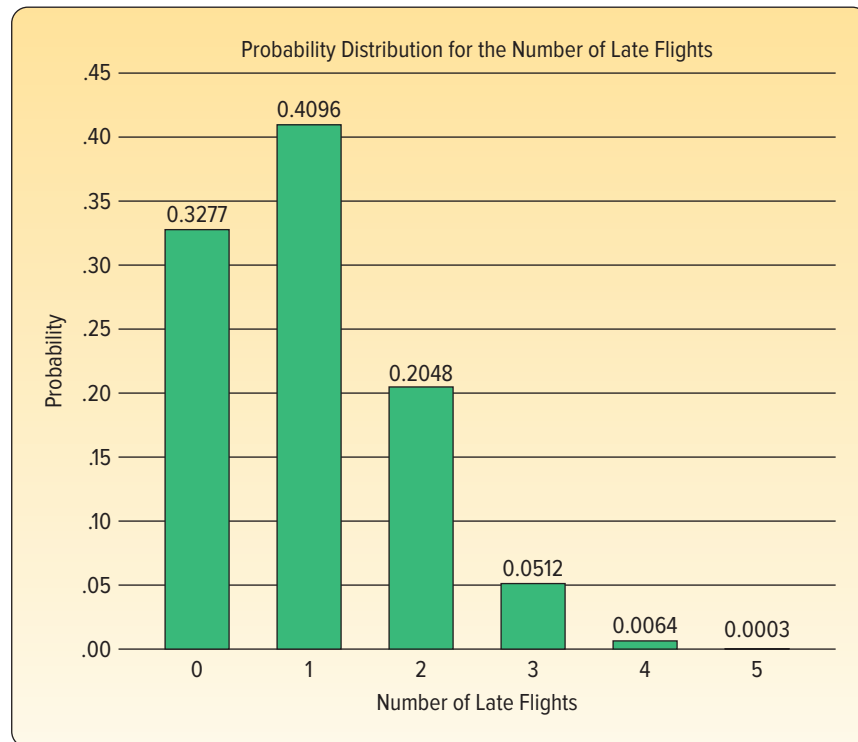
We can use formula (6–3). The probability that a particular flight is late is .20, so let  $\pi = .20$ . There are five flights, so  $n = 5$ , and  $X$ , the random variable, refers to the number of successes. In this case, a “success” is a flight that arrives late. The random variable,  $x$ , can be equal to 0 late flights in the five trials, 1 late flight in the five trials, or 2, 3, 4, or 5. The probability for no late arrivals,  $x = 0$ , is

$$\begin{aligned} P(0) &= {}_n C_x (\pi)^x (1 - \pi)^{n-x} \\ &= {}_5 C_0 (.20)^0 (1 - .20)^{5-0} = (1)(1)(.3277) = .3277 \end{aligned}$$

The probability that exactly one of the five flights will arrive late today is .4096, found by

$$\begin{aligned} P(1) &= {}_n C_x (\pi)^x (1 - \pi)^{n-x} \\ &= {}_5 C_1 (.20)^1 (1 - .20)^{5-1} = (5)(.20)(.4096) = .4096 \end{aligned}$$

The entire binomial probability distribution with  $\pi = .20$  and  $n = 5$  is shown in the following bar chart. We observe that the probability of exactly three late flights is .0512 and, from the bar chart, that the distribution of the number of late arrivals is positively skewed.



The mean ( $\mu$ ) and the variance ( $\sigma^2$ ) of a binomial distribution are computed in a “shortcut” fashion by:

**MEAN OF A BINOMIAL DISTRIBUTION**  $\mu = n\pi$  **(6-4)**

**VARIANCE OF A BINOMIAL DISTRIBUTION**  $\sigma^2 = n\pi(1 - \pi)$  **(6-5)**

For the example regarding the number of late flights, recall that  $\pi = .20$  and  $n = 5$ . Hence:

$$\mu = n\pi = (5)(.20) = 1.0$$

$$\sigma^2 = n\pi(1 - \pi) = 5(.20)(1 - .20) = .7997$$

The mean of 1.0 and the variance of .7997 can be verified from formulas (6-1) and (6-2). The probability distribution from the bar chart shown earlier and the details of the calculations are shown below.

| Number of Late Flights,<br>$x$ | $P(x)$ | $xP(x)$        | $x - \mu$ | $(x - \mu)^2$ | $(x - \mu)^2P(x)$   |
|--------------------------------|--------|----------------|-----------|---------------|---------------------|
| 0                              | 0.3277 | 0.0000         | -1        | 1             | 0.3277              |
| 1                              | 0.4096 | 0.4096         | 0         | 0             | 0                   |
| 2                              | 0.2048 | 0.4096         | 1         | 1             | 0.2048              |
| 3                              | 0.0512 | 0.1536         | 2         | 4             | 0.2048              |
| 4                              | 0.0064 | 0.0256         | 3         | 9             | 0.0576              |
| 5                              | 0.0003 | 0.0015         | 4         | 16            | 0.0048              |
|                                |        | $\mu = 1.0000$ |           |               | $\sigma^2 = 0.7997$ |

### Binomial Probability Tables

Formula (6-3) can be used to build a binomial probability distribution for any value of  $n$  and  $\pi$ . However, for a larger  $n$ , the calculations take more time. For convenience, the tables in Appendix B.1 show the result of using the formula for various values of  $n$  and  $\pi$ . Table 6-2 shows part of Appendix B.1 for  $n = 6$  and various values of  $\pi$ .

**TABLE 6-2** Binomial Probabilities for  $n = 6$  and Selected Values of  $\pi$

|                   |      | $n = 6$<br>Probability |      |      |      |      |      |      |      |      |      |
|-------------------|------|------------------------|------|------|------|------|------|------|------|------|------|
| $x \setminus \pi$ | .05  | .1                     | .2   | .3   | .4   | .5   | .6   | .7   | .8   | .9   | .95  |
| 0                 | .735 | .531                   | .262 | .118 | .047 | .016 | .004 | .001 | .000 | .000 | .000 |
| 1                 | .232 | .354                   | .393 | .303 | .187 | .094 | .037 | .010 | .002 | .000 | .000 |
| 2                 | .031 | .098                   | .246 | .324 | .311 | .234 | .138 | .060 | .015 | .001 | .000 |
| 3                 | .002 | .015                   | .082 | .185 | .276 | .313 | .276 | .185 | .082 | .015 | .002 |
| 4                 | .000 | .001                   | .015 | .060 | .138 | .234 | .311 | .324 | .246 | .098 | .031 |
| 5                 | .000 | .000                   | .002 | .010 | .037 | .094 | .187 | .303 | .393 | .531 | .232 |
| 6                 | .000 | .000                   | .000 | .001 | .004 | .016 | .047 | .118 | .262 | .531 | .735 |

**EXAMPLE**

In the Southwest, 5% of all cell phone calls are dropped. What is the probability that out of six randomly selected calls, none was dropped? Exactly one? Exactly two? Exactly three? Exactly four? Exactly five? Exactly six out of six?



**SOLUTION**

The binomial conditions are met: (a) there are only two possible outcomes (a particular call is either dropped or not dropped), (b) there are a fixed number of trials (6), (c) there is a constant probability of success (.05), and (d) the trials are independent.

Refer to Table 6–2 on the previous page for the probability of exactly zero dropped calls. Go down the left margin to an  $x$  of 0. Now move horizontally to the column headed by a  $\pi$  of .05 to find the probability. It is .735. The values in Table 6–2 are rounded to three decimal places.

The probability of exactly one dropped call in a sample of six calls is .232. The complete binomial probability distribution for  $n = 6$  and  $\pi = .05$  is:

| Number of Dropped Calls, $x$ | Probability of Occurrence, $P(x)$ | Number of Dropped Calls, $x$ | Probability of Occurrence, $P(x)$ |
|------------------------------|-----------------------------------|------------------------------|-----------------------------------|
| 0                            | .735                              | 4                            | .000                              |
| 1                            | .232                              | 5                            | .000                              |
| 2                            | .031                              | 6                            | .000                              |
| 3                            | .002                              |                              |                                   |

Of course, there is a slight chance of getting exactly five dropped calls out of six random selections. It is .00000178, found by inserting the appropriate values in the binomial formula:

$$P(5) = {}_6C_5(.50)^5(.95)^1 = (6)(.05)^5(.95) = .00000178$$

For six out of the six, the exact probability is .00000016. Thus, the probability is very small that five or six calls will be dropped in six trials.

We can compute the mean or expected value of the distribution of the number defective:

$$\mu = n\pi = (6)(.05) = 0.30$$

$$\sigma^2 = n\pi(1 - \pi) = 6(.05)(.95) = 0.285$$

Appendix B.1 is limited. It gives probabilities for  $n$  values from 1 to 15 and  $\pi$  values of .05, .10, . . . , .90, and .95. A software program can generate the probabilities for a specified number of successes, given  $n$  and  $\pi$ . The Excel output to the left shows the probability when  $n = 40$  and  $\pi = .09$ . Note that the number of successes stops at 15 because the probabilities for 16 to 40 are very close to 0. The instructions are detailed in the **Software Commands** in Appendix C.

Several additional points should be made regarding the binomial probability distribution.

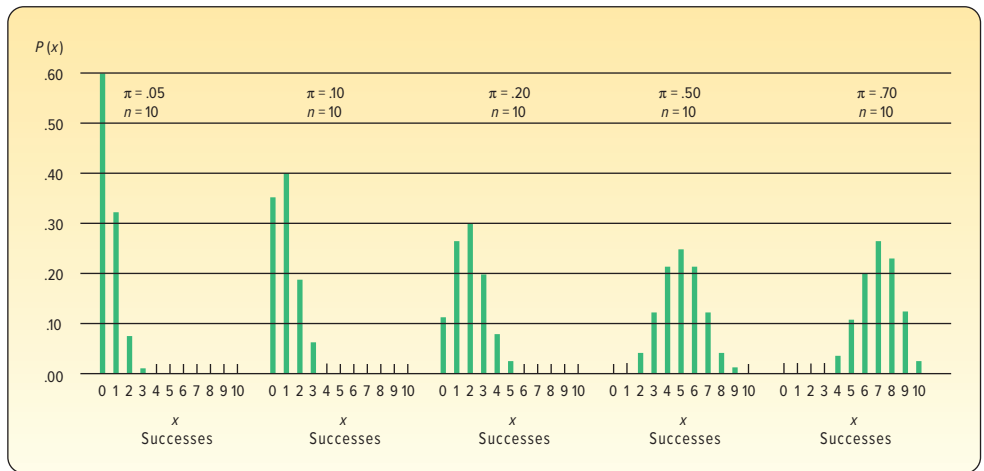
1. If  $n$  remains the same but  $\pi$  increases from .05 to .95, the shape of the distribution changes. Look at Table 6–3 and Chart 6–2 on the following page. The distribution for a  $\pi$  of .05 is positively skewed. As  $\pi$  approaches .50, the distribution becomes symmetrical. As  $\pi$  goes beyond .50 and moves toward .95, the probability distribution becomes negatively skewed. Table 6–3 highlights probabilities for  $n = 10$  and a  $\pi$  of .05, .10, .20, .50, and .70. The graphs of these probability distributions are shown in Chart 6–2.
2. If  $\pi$ , the probability of success, remains the same but  $n$  becomes larger, the shape of the binomial distribution becomes more symmetrical. Chart 6–3 on the following page shows a situation where  $\pi$  remains constant at .10 but  $n$  increases from 7 to 40.

|    | A       | B           |
|----|---------|-------------|
| 1  | Success | Probability |
| 2  | 0       | 0.0230      |
| 3  | 1       | 0.0910      |
| 4  | 2       | 0.1754      |
| 5  | 3       | 0.2198      |
| 6  | 4       | 0.2011      |
| 7  | 5       | 0.1432      |
| 8  | 6       | 0.0826      |
| 9  | 7       | 0.0397      |
| 10 | 8       | 0.0162      |
| 11 | 9       | 0.0057      |
| 12 | 10      | 0.0017      |
| 13 | 11      | 0.0005      |
| 14 | 12      | 0.0001      |
| 15 | 13      | 0.0000      |
| 16 | 14      | 0.0000      |
| 17 | 15      | 0.0000      |

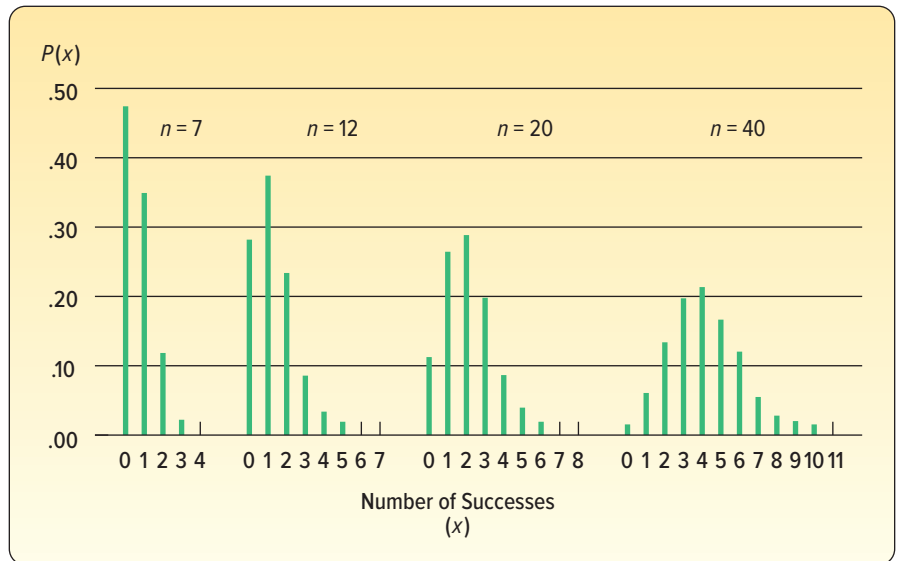
Source: Microsoft Excel

**TABLE 6-3** Probability of 0, 1, 2, . . . Successes for a  $\pi$  of .05, .10, .20, .50, and .70, and an  $n$  of 10

| $x \backslash \pi$ | .05  | .1   | .2   | .3   | .4   | .5   | .6   | .7   | .8   | .9   | .95  |
|--------------------|------|------|------|------|------|------|------|------|------|------|------|
| 0                  | .599 | .349 | .107 | .028 | .006 | .001 | .000 | .000 | .000 | .000 | .000 |
| 1                  | .315 | .387 | .268 | .121 | .040 | .010 | .002 | .000 | .000 | .000 | .000 |
| 2                  | .075 | .194 | .302 | .233 | .121 | .044 | .011 | .001 | .000 | .000 | .000 |
| 3                  | .010 | .057 | .201 | .267 | .215 | .117 | .042 | .009 | .001 | .000 | .000 |
| 4                  | .001 | .011 | .088 | .200 | .251 | .205 | .111 | .037 | .006 | .000 | .000 |
| 5                  | .000 | .001 | .026 | .103 | .201 | .246 | .201 | .103 | .026 | .001 | .000 |
| 6                  | .000 | .000 | .006 | .037 | .111 | .205 | .251 | .200 | .088 | .011 | .001 |
| 7                  | .000 | .000 | .001 | .009 | .042 | .117 | .215 | .267 | .201 | .057 | .010 |
| 8                  | .000 | .000 | .000 | .001 | .011 | .044 | .121 | .233 | .302 | .194 | .075 |
| 9                  | .000 | .000 | .000 | .000 | .002 | .010 | .040 | .121 | .268 | .387 | .315 |
| 10                 | .000 | .000 | .000 | .000 | .000 | .001 | .006 | .028 | .107 | .349 | .599 |



**CHART 6-2** Graphing the Binomial Probability Distribution for a  $\pi$  of .05, .10, .20, .50, and .70, and an  $n$  of 10



**CHART 6-3** Chart Representing the Binomial Probability Distribution for a  $\pi$  of .10 and an  $n$  of 7, 12, 20, and 40

## SELF-REVIEW 6-3



Ninety-five percent of the employees at the J. M. Smucker Company plant on Laskey Road have their bimonthly wages sent directly to their bank by electronic funds transfer. This is also called direct deposit. Suppose we select a random sample of seven employees.

- Does this situation fit the assumptions of the binomial distribution?
- What is the probability that all seven employees use direct deposit?
- Use formula (6-3) to determine the exact probability that four of the seven sampled employees use direct deposit.
- Use Excel to verify your answers to parts (b) and (c).

## EXERCISES

- In a binomial situation,  $n = 4$  and  $\pi = .25$ . Determine the probabilities of the following events using the binomial formula.
  - $x = 2$
  - $x = 3$
- In a binomial situation,  $n = 5$  and  $\pi = .40$ . Determine the probabilities of the following events using the binomial formula.
  - $x = 1$
  - $x = 2$
- Assume a binomial distribution where  $n = 3$  and  $\pi = .60$ .
  - Refer to Appendix B.1, and list the probabilities for values of  $x$  from 0 to 3.
  - Determine the mean and standard deviation of the distribution from the general definitions given in formulas (6-1) and (6-2).
- Assume a binomial distribution where  $n = 5$  and  $\pi = .30$ .
  - Refer to Appendix B.1 and list the probabilities for values of  $x$  from 0 to 5.
  - Determine the mean and standard deviation of the distribution from the general definitions given in formulas (6-1) and (6-2).
- An American Society of Investors survey found 30% of individual investors have used a discount broker. In a random sample of nine individuals, what is the probability:
  - Exactly two of the sampled individuals have used a discount broker?
  - Exactly four of them have used a discount broker?
  - None of them has used a discount broker?
- FILE** The U.S. Postal Service reports 95% of first-class mail within the same city is delivered within 2 days of the time of mailing. Six letters are randomly sent to different locations.
  - What is the probability that all six arrive within 2 days?
  - What is the probability that exactly five arrive within 2 days?
  - Find the mean number of letters that will arrive within 2 days.
  - Compute the variance and standard deviation of the number that will arrive within 2 days.
- FILE** Industry standards suggest that 10% of new vehicles require warranty service within the first year. Jones Nissan in Sumter, South Carolina, sold 12 Nissans yesterday.
  - What is the probability that none of these vehicles requires warranty service?
  - What is the probability exactly one of these vehicles requires warranty service?
  - Determine the probability that exactly two of these vehicles require warranty service.
  - Compute the mean and standard deviation of this probability distribution.
- FILE** A telemarketer makes six phone calls per hour and is able to make a sale on 30% of these contacts. During the next 2 hours, find:
  - The probability of making exactly four sales.
  - The probability of making no sales.
  - The probability of making exactly two sales.
  - The mean number of sales in the 2-hour period.

17. **FILE** A recent survey by the American Accounting Association revealed 23% of students graduating with a major in accounting select public accounting. Suppose we select a sample of 15 recent graduates.
- What is the probability two select public accounting?
  - What is the probability five select public accounting?
  - How many graduates would you expect to select public accounting?
18. **FILE** It is reported that 41% of American households use a cell phone exclusively for their telephone service. In a sample of eight households:
- Find the probability that no household uses a cell phone as their exclusive telephone service.
  - Find the probability that exactly 5 households exclusively use a cell phone for telephone service.
  - Find the mean number of households exclusively using cell phones.

## Cumulative Binomial Probability Distributions

We may wish to know the probability of correctly guessing the answers to 6 or more true/false questions out of 10. Or we may be interested in the probability of selecting less than two defectives at random from production during the previous hour. In these cases, we need cumulative frequency distributions similar to the ones developed in the Chapter 2, Cumulative Distributions section on page 39. The following example will illustrate.

### EXAMPLE

A study by the Illinois Department of Transportation concluded that 76.2% of front seat occupants used seat belts. That is, both occupants of the front seat were using their seat belts. Suppose we decide to compare that information with current usage. We select a sample of 12 vehicles.

- What is the probability the front seat occupants in exactly 7 of the 12 vehicles selected are wearing seat belts?
- What is the probability the front seat occupants in at least 7 of the 12 vehicles are wearing seat belts?

### SOLUTION

This situation meets the binomial requirements.

- In a particular vehicle, both the front seat occupants are either wearing seat belts or they are not. There are only two possible outcomes.
- There are a fixed number of trials, 12 in this case, because 12 vehicles are checked.
- The probability of a “success” (occupants wearing seat belts) is the same from one vehicle to the next: 76.2%.
- The trials are independent. If the fourth vehicle selected in the sample has all the occupants wearing their seat belts, this does not have any effect on the results for the fifth or tenth vehicle.

To find the likelihood the occupants of exactly 7 of the sampled vehicles are wearing seat belts, we use formula (6–3). In this case,  $n = 12$  and  $\pi = .762$ .

$$P(x = 7) = {}_{12}C_7(.762)^7(1 - .762)^{12-7} = 792(.149171)(.000764) = .0902$$

So we conclude the likelihood that the occupants of exactly 7 of the 12 sampled vehicles will be wearing their seat belts is about 9%.

To find the probability that the occupants in 7 or more of the vehicles will be wearing seat belts, we use formula (6–3) from this chapter as well as the special rule of addition from the previous chapter. See formula (5-2) on page 126.

Because the events are mutually exclusive (meaning that a particular sample of 12 vehicles cannot have both a *total* of 7 and a *total* of 8 vehicles where the occupants are wearing seat belts), we find the probability of 7 vehicles where the occupants are wearing seat belts, the probability of 8, and so on up to the probability that occupants of all 12 sample vehicles are wearing seat belts. The probability of each of these outcomes is then totaled.

$$\begin{aligned} P(x \geq 7) &= P(x = 7) + P(x = 8) + P(x = 9) + P(x = 10) + P(x = 11) + P(x = 12) \\ &= .0902 + .1805 + .2569 + .2467 + .1436 + .0383 \\ &= .9562 \end{aligned}$$

So the probability of selecting 12 cars and finding that the occupants of 7 or more vehicles were wearing seat belts is .9562. This information is shown on the following Excel spreadsheet. There is a slight difference in the software answer due to rounding. The Excel commands are similar to those detailed in the **Software Commands** in Appendix C.

|    | A       | B           | C | D |
|----|---------|-------------|---|---|
| 1  | Success | Probability |   |   |
| 2  | 0       | 0.0000      |   |   |
| 3  | 1       | 0.0000      |   |   |
| 4  | 2       | 0.0000      |   |   |
| 5  | 3       | 0.0002      |   |   |
| 6  | 4       | 0.0017      |   |   |
| 7  | 5       | 0.0088      |   |   |
| 8  | 6       | 0.0329      |   |   |
| 9  | 7       | 0.0902      |   |   |
| 10 | 8       | 0.1805      |   |   |
| 11 | 9       | 0.2569      |   |   |
| 12 | 10      | 0.2467      |   |   |
| 13 | 11      | 0.1436      |   |   |
| 14 | 12      | 0.0383      |   |   |
| 15 |         | 0.9563      |   |   |

Source: Microsoft Excel

## SELF-REVIEW 6-4



A recent study revealed that 40% of women in the San Diego metropolitan area who work full time also volunteer in the community. Suppose we randomly select eight women in the San Diego area.

- What are the values for  $n$  and  $\pi$ ?
- What is the probability exactly three of the women volunteer in the community?
- What is the probability at least one of the women volunteers in the community?

## EXERCISES

- In a binomial distribution,  $n = 8$  and  $\pi = .30$ . Find the probabilities of the following events.
  - $x = 2$ .
  - $x \leq 2$  (the probability that  $x$  is equal to or less than 2).
  - $x \geq 3$  (the probability that  $x$  is equal to or greater than 3).

20. In a binomial distribution,  $n = 12$  and  $\pi = .60$ . Find the following probabilities.
- $x = 5$ .
  - $x \leq 5$ .
  - $x \geq 6$ .
21. **FILE** In a recent study, 90% of the homes in the United States were found to have large-screen TVs. In a sample of nine homes, what is the probability that:
- All nine have large-screen TVs?
  - Less than five have large-screen TVs?
  - More than five have large-screen TVs?
  - At least seven homes have large-screen TVs?
22. **FILE** A manufacturer of window frames knows from long experience that 5% of the production will have some type of minor defect that will require an adjustment. What is the probability that in a sample of 20 window frames:
- None will need adjustment?
  - At least one will need adjustment?
  - More than two will need adjustment?
23. **FILE** The speed with which utility companies can resolve problems is very important. GTC, the Georgetown Telephone Company, reports it can resolve customer problems the same day they are reported in 70% of the cases. Suppose the 15 cases reported today are representative of all complaints.
- How many of the problems would you expect to be resolved today? What is the standard deviation?
  - What is the probability 10 of the problems can be resolved today?
  - What is the probability 10 or 11 of the problems can be resolved today?
  - What is the probability more than 10 of the problems can be resolved today?
24. **FILE** It is asserted that 80% of the cars approaching an individual toll booth in New Jersey are equipped with an E-ZPass transponder. Find the probability that in a sample of six cars:
- All six will have the transponder.
  - At least three will have the transponder.
  - None will have a transponder.

**LO6-5**

Explain the assumptions of the Poisson distribution and apply it to calculate probabilities.

## POISSON PROBABILITY DISTRIBUTION

The Poisson probability distribution is a discrete distribution that describes the probability of outcomes from a Poisson experiment. In a **Poisson experiment** we count the number of times some event occurs during a specified interval. Examples of an interval may be time, distance, area, or volume. A Poisson probability experiment must have these characteristics:

### POISSON EXPERIMENT

- The random variable is the number of times some event occurs during a defined interval.
- The probability of the event is proportional to the size of the interval.
- The intervals do not overlap and are independent.

A Poisson experiment is based on two requirements. The first requirement is that the probability of an event is proportional to the length of the interval. The second requirement is that the intervals are independent. To put it another way, the longer the interval, the higher the probability of an outcome, and the number of occurrences in one interval does not affect the other intervals.

When an experiment is consistent with the Poisson requirements, the Poisson probability distribution is used to calculate the probability of experimental outcomes. The Poisson probability distribution is a limiting form of the binomial distribution when the

**STATISTICS IN ACTION**

Near the end of World War II, the Germans developed rocket bombs, which were fired at the city of London. The Allied military command didn't know whether these bombs were fired at random or whether they had an aiming device. To investigate, the city of London was divided into 586 square regions. The distribution of hits in each square was recorded as follows:

|         |     |     |    |    |   |   |
|---------|-----|-----|----|----|---|---|
| Hits    | 0   | 1   | 2  | 3  | 4 | 5 |
| Regions | 229 | 221 | 93 | 35 | 7 | 1 |

To interpret, the above chart indicates that 229 regions were not hit with one of the bombs. Seven regions were hit four times. Using the Poisson distribution, with a mean of 0.93 hits per region, the expected number of hits is as follows:

|         |       |       |       |      |     |           |
|---------|-------|-------|-------|------|-----|-----------|
| Hits    | 0     | 1     | 2     | 3    | 4   | 5 or more |
| Regions | 231.2 | 215.0 | 100.0 | 31.0 | 7.2 | 1.6       |

Because the actual number of hits was close to the expected number of hits, the military command concluded that the bombs were falling at random. The Germans had not developed a bomb with an aiming device.

probability of a success is very small and  $n$  is large. It is often referred to as the “law of improbable events,” meaning that the probability,  $\pi$ , of a particular event’s happening is quite small.

This probability distribution has many applications. It is used as a model to describe the distribution of errors in data entry, the number of scratches and other imperfections in newly painted car panels, the number of defective parts in outgoing shipments, the number of customers waiting to be served at a restaurant or waiting to get into an attraction at Disney World, and the number of accidents on I–75 during a three-month period.

The **Poisson probability distribution** is described mathematically by the formula:

$$\text{POISSON PROBABILITY DISTRIBUTION} \quad P(x) = \frac{\mu^x e^{-\mu}}{x!} \quad (6-6)$$

where:

$\mu$  (mu) is the mean number of occurrences (successes) in a particular interval.

$e$  is the constant 2.71828 (base of the Napierian logarithmic system).

$x$  is the number of occurrences (successes).

$P(x)$  is the probability for a specified value of  $x$ .

The mean number of successes,  $\mu$ , is found by  $n\pi$ , where  $n$  is the total number of trials and  $\pi$  the probability of success.

$$\text{MEAN OF A POISSON DISTRIBUTION} \quad \mu = n\pi \quad (6-7)$$

The variance of the Poisson is equal to its mean. If, for example, the probability that a check cashed by a bank will bounce is .0003, and 10,000 checks are cashed, the mean and the variance for the number of bad checks is 3.0, found by  $\mu = n\pi = 10,000(.0003) = 3.0$ .

Recall that for a binomial distribution there are a fixed number of trials. For example, for a four-question multiple-choice test there can only be zero, one, two, three, or four successes (correct answers). The random variable  $x$  for a Poisson distribution, however, can assume an *infinite number of values*—that is, 0, 1, 2, 3, 4, 5, . . . However, *the probabilities become very small after the first few occurrences* (successes).

**EXAMPLE**

Budget Airlines is a seasonal airline that operates flights from Myrtle Beach, South Carolina, to various cities in the Northeast. The destinations include Boston, Pittsburgh, Buffalo, and both LaGuardia and JFK airports in New York City. Recently, Budget has been concerned about the number of lost bags. Ann Poston from the Analytics Department was asked to study the issue. She randomly selected a sample of 500 flights and found that a total of 20 bags were lost on the sampled flights.

Show that this situation is a Poisson experiment. What is the mean number of bags lost per flight? What is the likelihood that no bags are lost on a flight? What is the probability at least one bag is lost?

**SOLUTION**

To begin, let’s confirm that the Budget Airlines situation follows a Poisson experiment. Refer to the highlighted box labeled Poisson Experiment in this section. We count the number of bags lost on a particular flight. So, the random variable is the number of lost bags. A flight is the interval. We can make an assumption that the probability of a lost bag increases with the length of a flight. Lastly, the number of bags lost on a flight is independent of the number of bags lost on other flights.





Earlier in this section, we mentioned that the Poisson probability distribution is a limiting form of the binomial. That is, we could estimate a binomial probability using the Poisson. In the following example, we use the Poisson distribution to estimate a binomial probability when  $n$ , the number of trials, is large and  $\pi$ , the probability of a success, is small.

### EXAMPLE

Coastal Insurance Company underwrites insurance for beachfront properties along the Virginia, North Carolina, South Carolina, and Georgia coasts. It uses the estimate that the probability of a named Category III hurricane (sustained winds of more than 110 miles per hour) or higher striking a particular region of the coast (for example, St. Simons Island, Georgia) in any one year is .05. If a homeowner takes a 30-year mortgage on a recently purchased property in St. Simons, what is the likelihood that the owner will experience at least one hurricane during the mortgage period?

### SOLUTION

To use the Poisson probability distribution, we begin by determining the mean or expected number of storms meeting the criteria hitting St. Simons during the 30-year period. That is:

$$\mu = n\pi = 30(.05) = 1.5$$

where:

$n$  is the number of years, 30 in this case.

$\pi$  is the probability a hurricane meeting the strength criteria comes ashore.

$\mu$  is the mean or expected number of storms in a 30-year period.

To find the probability of at least one storm hitting St. Simons Island, Georgia, we first find the probability of no storms hitting the coast and subtract that value from 1.

$$P(x \geq 1) = 1 - P(x = 0) = 1 - \frac{\mu^0 e^{-1.5}}{0!} = 1 - .2231 = .7769$$

We conclude that the likelihood a hurricane meeting the strength criteria will strike the beachfront property at St. Simons during the 30-year period when the mortgage is in effect is .7769. To put it another way, the probability St. Simons will be hit by a Category III or higher hurricane during the 30-year period is a little more than 75%.

We should emphasize that the continuum, as previously described, still exists. That is, there are expected to be 1.5 storms hitting the coast per 30-year period. The continuum is the 30-year period.

In the preceding case, we are actually using the Poisson distribution as an estimate of the binomial. Note that we've met the binomial conditions outlined on page 165.

- There are only two possible outcomes: a hurricane hits the St. Simons area or it does not.
- There are a fixed number of trials, in this case 30 years.
- There is a constant probability of success; that is, the probability of a hurricane hitting the area is .05 each year.
- The years are independent. That means if a named storm strikes in the fifth year, that has no effect on any other year.

To find the probability of at least one storm striking the area in a 30-year period using the binomial distribution:

$$P(x \geq 1) = 1 - P(x = 0) = 1 - [{}_{30}C_0(.05)^0(.95)^{30}] = 1 - [(1)(1)(.2146)] = .7854$$

The probability of at least one hurricane hitting the St. Simons area during the 30-year period using the binomial distribution is .7854.

Which answer is correct? Why should we look at the problem both ways? The binomial is the more “technically correct” solution. The Poisson can be thought of as an approximation for the binomial, when  $n$ , the number of trials, is large, and  $\pi$ , the probability of a success, is small. We look at the problem using both distributions to emphasize the convergence of the two discrete distributions. In some instances, using the Poisson may be the quicker solution, and, as you see, there is little practical difference in the answers. In fact, as  $n$  gets larger and  $\pi$  smaller, the difference between the two distributions gets smaller.

The Poisson probability distribution is always positively skewed and the random variable has no specific upper limit. In the lost bags example/solution, the Poisson distribution, with  $\mu = 0.04$ , is highly skewed. As  $\mu$  becomes larger, the Poisson distribution becomes more symmetrical. For example, Chart 6–4 shows the distributions of the number of transmission services, muffler replacements, and oil changes per day at Avellino’s Auto Shop. They follow Poisson distributions with means of 0.7, 2.0, and 6.0, respectively.

In summary, the Poisson distribution is a family of discrete distributions. All that is needed to construct a Poisson probability distribution is the mean number of defects, errors, or other random variable, designated as  $\mu$ .

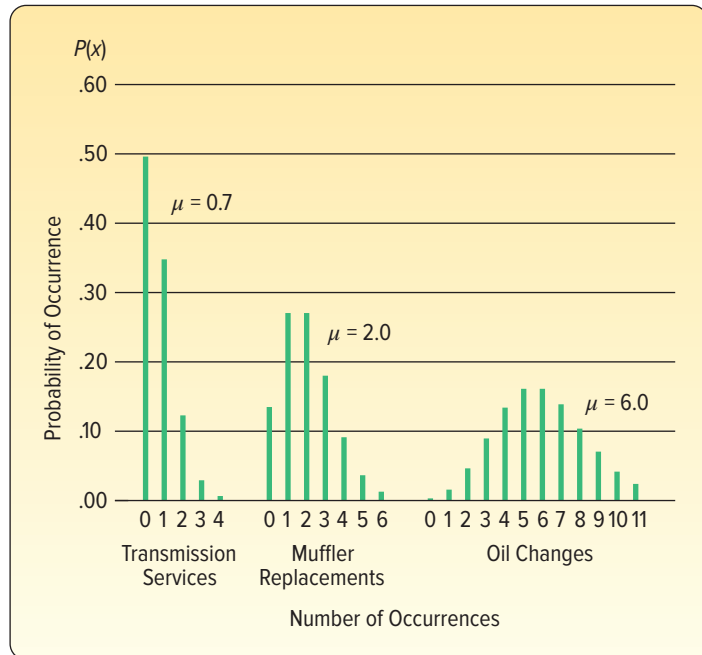


CHART 6–4 Poisson Probability Distributions for Means of 0.7, 2.0, and 6.0

## SELF-REVIEW 6–5



From actuary tables, Washington Insurance Company determined the likelihood that a man age 25 will die within the next year is .0002. If Washington Insurance sells 4,000 policies to 25-year-old men this year, what is the probability they will pay on exactly one policy?

## EXERCISES

25. In a Poisson distribution,  $\mu = 0.4$ .
  - a. What is the probability that  $x = 0$ ?
  - b. What is the probability that  $x > 0$ ?
26. In a Poisson distribution,  $\mu = 4$ .
  - a. What is the probability that  $x = 2$ ?
  - b. What is the probability that  $x \leq 2$ ?
  - c. What is the probability that  $x > 2$ ?
27. Ms. Bergen is a loan officer at Coast Bank and Trust. From her years of experience, she estimates that the probability is .025 that an applicant will not be able to repay his or her installment loan. Last month she made 40 loans.
  - a. What is the probability that three loans will be defaulted?
  - b. What is the probability that at least three loans will be defaulted?
28. Automobiles arrive at the Elkhart exit of the Indiana Toll Road at the rate of two per minute. The distribution of arrivals approximates a Poisson distribution.
  - a. What is the probability that no automobiles arrive in a particular minute?
  - b. What is the probability that at least one automobile arrives during a particular minute?
29. It is estimated that 0.5% of the callers to the Customer Service department of Dell Inc. will receive a busy signal. What is the probability that of today's 1,200 callers, at least 5 received a busy signal?
30. In the past, schools in Los Angeles County have closed an average of 3 days each year for weather emergencies. What is the probability that schools in Los Angeles County will close for 4 days next year?

## CHAPTER SUMMARY

- I. A random variable is a numerical value determined by the outcome of an experiment.
- II. A probability distribution is a listing of all possible outcomes of an experiment and the probability associated with each outcome.
  - A. A discrete probability distribution can assume only certain values. The main features are:
    1. The sum of the probabilities is 1.00.
    2. The probability of a particular outcome is between 0.00 and 1.00.
    3. The outcomes are mutually exclusive.
  - B. A continuous distribution can assume an infinite number of values within a specific range.
- III. The mean and variance of a probability distribution are computed as follows.
  - A. The mean is equal to:
 
$$\mu = \sum[xP(x)] \quad (6-1)$$
  - B. The variance is equal to:
 
$$\sigma^2 = \sum[(x - \mu)^2P(x)] \quad (6-2)$$
- IV. The binomial distribution has the following characteristics.
  - A. Each outcome is classified into one of two mutually exclusive categories.
  - B. The distribution results from a count of the number of successes in a fixed number of trials.
  - C. The probability of a success remains the same from trial to trial.
  - D. Each trial is independent.
  - E. A binomial probability is determined as follows:
 
$$P(x) = {}_n C_x \pi^x (1 - \pi)^{n-x} \quad (6-3)$$

F. The mean is computed as:

$$\mu = n\pi \quad (6-4)$$

G. The variance is:

$$\sigma^2 = n\pi(1 - \pi) \quad (6-5)$$

V. The Poisson distribution has the following characteristics.

- A. It describes the number of times some event occurs during a specified interval.
- B. The probability of a “success” is proportional to the length of the interval.
- C. Nonoverlapping intervals are independent.
- D. It is a limiting form of the binomial distribution when  $n$  is large and  $\pi$  is small.
- E. A Poisson probability is determined from the following equation:

$$P(x) = \frac{\mu^x e^{-\mu}}{x!} \quad (6-6)$$

F. The mean and the variance are:

$$\begin{aligned} \mu &= n\pi \\ \sigma^2 &= n\pi \end{aligned} \quad (6-7)$$

## CHAPTER EXERCISES

31. What is the difference between a random variable and a probability distribution?
32. For each of the following, indicate whether the random variable is discrete or continuous.
  - a. The length of time to get a haircut.
  - b. The number of cars a jogger passes each morning while running.
  - c. The number of hits for a team in a high school girls' softball game.
  - d. The number of patients treated at the South Strand Medical Center between 6 and 10 p.m. each night.
  - e. The distance your car traveled on the last fill-up.
  - f. The number of customers at the Oak Street Wendy's who used the drive-through facility.
  - g. The distance between Gainesville, Florida, and all Florida cities with a population of at least 50,000.
33. **FILE** An investment will be worth \$1,000, \$2,000, or \$5,000 at the end of the year. The probabilities of these values are .25, .60, and .15, respectively. Determine the mean and variance of the investment's dollar value.
34. The following notice appeared in the golf shop at a Myrtle Beach, South Carolina, golf course.

### Blackmoor Golf Club Members

The golf shop is holding a raffle to win a TaylorMade M1 10.5° Regular Flex Driver (\$300 value).

Tickets are \$5.00 each.

Only 80 tickets will be sold.

Please see the golf shop to get your tickets!

John Underpar buys a ticket.

- a. What are Mr. Underpar's possible monetary outcomes?
- b. What are the probabilities of the possible outcomes?
- c. Summarize Mr. Underpar's “experiment” as a probability distribution.
- d. What is the mean or expected value of the probability distribution? Explain your result.
- e. If all 80 tickets are sold, what is the expected return to the Club?

35. **FILE** Croissant Bakery Inc. offers special decorated cakes for birthdays, weddings, and other occasions. It also has regular cakes available in its bakery. The following table gives the total number of cakes sold per day and the corresponding probability. Compute the mean, variance, and standard deviation of the number of cakes sold per day.

| Number of Cakes Sold in a Day | Probability |
|-------------------------------|-------------|
| 12                            | .25         |
| 13                            | .40         |
| 14                            | .25         |
| 15                            | .10         |

36. **FILE** The payouts for the Powerball lottery and their corresponding odds and probabilities of occurrence are shown below. The price of a ticket is \$1.00. Find the mean and standard deviation of the payout. Hint: Don't forget to include the cost of the ticket and its corresponding probability.

| Divisions            | Payout       | Odds        | Probability    |
|----------------------|--------------|-------------|----------------|
| Five plus Powerball  | \$50,000,000 | 146,107,962 | 0.00000006844  |
| Match 5              | 200,000      | 3,563,609   | 0.00000280614  |
| Four plus Powerball  | 10,000       | 584,432     | 0.000001711060 |
| Match 4              | 100          | 14,255      | 0.000070145903 |
| Three plus Powerball | 100          | 11,927      | 0.000083836351 |
| Match 3              | 7            | 291         | 0.003424657534 |
| Two plus Powerball   | 7            | 745         | 0.001340482574 |
| One plus Powerball   | 4            | 127         | 0.007812500000 |
| Zero plus Powerball  | 3            | 69          | 0.014285714286 |

37. In a recent study, 35% of people surveyed indicated chocolate was their favorite flavor of ice cream. Suppose we select a sample of 10 people and ask them to name their favorite flavor of ice cream.
- How many of those in the sample would you expect to name chocolate?
  - What is the probability exactly four of those in the sample name chocolate?
  - What is the probability four or more name chocolate?
38. **FILE** Thirty percent of the population in a Southwestern community are Spanish-speaking Americans. A Spanish-speaking person is accused of killing a non-Spanish-speaking American and goes to trial. Of the first 12 potential jurors, only 2 are Spanish-speaking Americans, and 10 are not. The defendant's lawyer challenges the jury selection, claiming bias against her client. The government lawyer disagrees, saying that the probability of this particular jury composition is common. Compute the probability and discuss the assumptions.
39. An auditor for Health Maintenance Services of Georgia reports 40% of policyholders 55 years or older submit a claim during the year. Fifteen policyholders are randomly selected for company records.
- How many of the policyholders would you expect to have filed a claim within the last year?
  - What is the probability that 10 of the selected policyholders submitted a claim last year?
  - What is the probability that 10 or more of the selected policyholders submitted a claim last year?
  - What is the probability that more than 10 of the selected policyholders submitted a claim last year?
40. Tire and Auto Supply is considering a 2-for-1 stock split. Before the transaction is finalized, at least two-thirds of the 1,200 company stockholders must approve the proposal. To evaluate the likelihood the proposal will be approved, the CFO selected a sample of 18 stockholders. He contacted each and found 14 approved of the proposed split. What is the likelihood of this event, assuming two-thirds of the stockholders approve?

41. A federal study reported that 7.5% of the U.S. workforce has a drug problem. A drug enforcement official for the state of Indiana wished to investigate this statement. In her sample of 20 employed workers:
- How many would you expect to have a drug problem? What is the standard deviation?
  - What is the likelihood that *none* of the workers sampled has a drug problem?
  - What is the likelihood *at least one* has a drug problem?
42. The Bank of Hawaii reports that 7% of its credit card holders will default at some time in their life. The Hilo branch just mailed out 12 new cards today.
- How many of these new cardholders would you expect to default? What is the standard deviation?
  - What is the likelihood that *none* of the cardholders will default?
  - What is the likelihood *at least one* will default?
43. Recent statistics suggest that 15% of those who visit a retail site on the Internet make a purchase. A retailer wished to verify this claim. To do so, she selected a sample of 16 “hits” to her site and found that 4 had actually made a purchase.
- What is the likelihood of exactly four purchases?
  - How many purchases should she expect?
  - What is the likelihood that four or more “hits” result in a purchase?
44. Acceptance sampling is a statistical method used to monitor the quality of purchased parts and components. To ensure the quality of incoming parts, a purchaser or manufacturer normally samples 20 parts and allows one defect.
- What is the likelihood of accepting a lot that is 1% defective?
  - If the quality of the incoming lot was actually 2%, what is the likelihood of accepting it?
  - If the quality of the incoming lot was actually 5%, what is the likelihood of accepting it?
45. Unilever Inc. recently developed a new body wash with a scent of ginger. Their research indicates that 30% of men like the new scent. To further investigate, Unilever’s marketing research group randomly selected 15 men and asked them if they liked the scent. What is the probability that six or more men like the ginger scent in the body wash?
46. Dr. Richmond, a psychologist, is studying the daytime television viewing habits of college students. She believes 45% of college students watch soap operas during the afternoon. To further investigate, she selects a sample of 10.
- Develop a probability distribution for the number of students in the sample who watch soap operas.
  - Find the mean and the standard deviation of this distribution.
  - What is the probability of finding exactly four students who watch soap operas?
  - What is the probability less than half of the students selected watch soap operas?
47. **FILE** A recent study conducted by Penn, Shone, and Borland, on behalf of LastMinute.com, revealed that 52% of business travelers plan their trips less than two weeks before departure. The study is to be replicated in the tri-state area with a sample of 12 frequent business travelers.
- Develop a probability distribution for the number of travelers who plan their trips within two weeks of departure.
  - Find the mean and the standard deviation of this distribution.
  - What is the probability exactly 5 of the 12 selected business travelers plan their trips within two weeks of departure?
  - What is the probability 5 or fewer of the 12 selected business travelers plan their trips within two weeks of departure?
48. Suppose 1.5% of the antennas on new Nokia cell phones are defective. For a random sample of 200 antennas, find the probability that:
- None of the antennas is defective.
  - Three or more of the antennas are defective.
49. A study of the checkout lines at the Safeway Supermarket in the South Strand area revealed that between 4 and 7 p.m. on weekdays there is an average of four customers waiting in line. What is the probability that you visit Safeway today during this period and find:
- No customers are waiting?
  - Four customers are waiting?
  - Four or fewer are waiting?
  - Four or more are waiting?

- 50.** An internal study by the Technology Services department at Lahey Electronics revealed company employees receive an average of two non-work-related e-mails per hour. Assume the arrival of these e-mails is approximated by the Poisson distribution.
- What is the probability Linda Lahey, company president, received exactly one non-work-related e-mail between 4 p.m. and 5 p.m. yesterday?
  - What is the probability she received five or more non-work-related e-mails during the same period?
  - What is the probability she did not receive any non-work-related e-mails during the period?
- 51.** Recent crime reports indicate that 3.1 motor vehicle thefts occur each minute in the United States. Assume that the distribution of thefts per minute can be approximated by the Poisson probability distribution.
- Calculate the probability exactly *four* thefts occur in a minute.
  - What is the probability there are *no* thefts in a minute?
  - What is the probability there is *at least one* theft in a minute?
- 52.** Recent difficult economic times have caused an increase in the foreclosure rate of home mortgages. Statistics from the Penn Bank and Trust Company show their monthly foreclosure rate is now 1 loan out of every 136 loans. Last month the bank approved 300 loans.
- How many foreclosures would you expect the bank to have last month?
  - What is the probability of exactly two foreclosures?
  - What is the probability of at least one foreclosure?
- 53.** The National Aeronautics and Space Administration (NASA) has experienced two disasters. The *Challenger* exploded over the Atlantic Ocean in 1986, and the *Columbia* disintegrated on reentry over East Texas in 2003. Based on the first 113 missions, and assuming failures occur at the same rate, consider the next 23 missions. What is the probability of exactly two failures? What is the probability of no failures?
- 54.** According to the “January theory,” if the stock market is up for the month of January, it will be up for the year. If it is down in January, it will be down for the year. According to an article in *The Wall Street Journal*, this theory held for 29 out of the last 34 years. Suppose there is no truth to this theory; that is, the probability it is either up or down is .50. What is the probability this could occur by chance? You will probably need a software package such as Excel or Minitab.
- 55.** During the second round of the 1989 U.S. Open golf tournament, four golfers scored a hole in one on the sixth hole. The odds of a professional golfer making a hole in one are estimated to be 3,708 to 1, so the probability is  $1/3,709$ . There were 155 golfers participating in the second round that day. Estimate the probability that four golfers would score a hole in one on the sixth hole.
- 56.** According to sales information in the first quarter of 2016, 2.7% of new vehicles sold in the United States were hybrids. This is down from 3.3% for the same period a year earlier. An analyst’s review of the data indicates that the reasons for the sales decline include the low price of gasoline and the higher price of a hybrid compared to similar vehicles. Let’s assume these statistics remain the same for 2017. That is, 2.7% of new car sales are hybrids in the first quarter of 2017. For a sample of 40 vehicles sold in the Richmond, Virginia, area:
- How many vehicles would you expect to be hybrid?
  - Use the Poisson distribution to find the probability that five of the sales were hybrid vehicles.
  - Use the binomial distribution to find the probability that five of the sales were hybrid vehicles.
- 57.** A recent CBS News survey reported that 67% of adults felt the U.S. Treasury should continue making pennies. Suppose we select a sample of 15 adults.
- How many of the 15 would we expect to indicate that the Treasury should continue making pennies? What is the standard deviation?
  - What is the likelihood that exactly eight adults would indicate the Treasury should continue making pennies?
  - What is the likelihood that at least eight adults would indicate the Treasury should continue making pennies?

## DATA ANALYTICS

(The data for these exercises are available in Connect.)

58. **FILE** Refer to the North Valley Real Estate data, which report information on homes sold in the area last year.
- Create a probability distribution for the number of bedrooms. Compute the mean and the standard deviation of this distribution.
  - Create a probability distribution for the number of bathrooms. Compute the mean and the standard deviation of this distribution.
59. **FILE** Refer to the Baseball 2016 data. Compute the mean number of home runs per game. To do this, first find the mean number of home runs per team for 2016. Next, divide this value by 162 (a season comprises 162 games). Then multiply by 2 because there are two teams in each game. Use the Poisson distribution to estimate the number of home runs that will be hit in a game. Find the probability that:
- There are no home runs in a game.
  - There are two home runs in a game.
  - There are at least four home runs in a game.

## PRACTICE TEST

## Part 1—Objective

- A listing of the possible outcomes of an experiment and the probability associated with each outcome is called a \_\_\_\_\_.
- The essential difference between a discrete random variable and a discrete probability distribution is that a discrete probability distribution includes the \_\_\_\_\_.
- In a discrete probability distribution, the sum of the possible probabilities is always equal to \_\_\_\_\_.
- The expected value of a probability distribution is also called the \_\_\_\_\_.
- How many outcomes are there in a particular binomial trial? \_\_\_\_\_
- Under what conditions will the probability of a success change from trial to trial in a binomial experiment? \_\_\_\_\_
- In a Poisson experiment, the mean and variance are \_\_\_\_\_.
- The Poisson distribution is a limiting case of the binomial probability distribution when  $n$  is large and \_\_\_\_\_ is small.
- Suppose 5% of patients who take a certain drug suffer undesirable side effects. If we select 10 patients currently taking the drug, what is the probability exactly two suffer undesirable side effects? \_\_\_\_\_
- The mean number of work-related accidents per month in a manufacturing plant is 1.70. What is the probability there will be no work-related accidents in a particular month? \_\_\_\_\_

## Part 2—Problems

- IRS data show that 15% of personal tax returns reporting an adjusted gross income more than \$1,000,000 will be subject to a computer audit. This year a CPA completed 16 returns with adjusted gross incomes more than \$1,000,000. The CPA wants to know the likelihoods that the returns will be audited.
  - What probability distribution applies to this situation?
  - What is the probability exactly one of these returns is audited?
  - What is the probability at least one of these returns is audited?
- For certain personal tax returns, the IRS will compute the amount to refund a taxpayer. Suppose the Cincinnati office of the IRS processes an average of three returns per hour that require a refund calculation.
  - What probability distribution applies to this situation?
  - What is the probability the IRS processes exactly three returns in a particular hour that require a refund calculation?
  - What is the probability the IRS does not compute a refund on any return in an hour?
  - What is the probability the IRS processes at least one return in a particular hour that requires a refund calculation?
- A CPA studied the number of exemptions claimed on tax returns. The data are summarized in the following table.

| Exemptions | Percent |
|------------|---------|
| 1          | 20%     |
| 2          | 50      |
| 3          | 20      |
| 4          | 10      |

- What is the mean number of exemptions claimed?
- What is the variance of the number of exemptions claimed?



# 7

# Continuous Probability Distributions



©Barry Austin Photography/Getty Images

- ▲ **MOST FOUR-YEAR** automobile leases allow up to 60,000 miles. If the lessee goes beyond this amount, a penalty of 20 cents per mile is added to the lease cost. Suppose the distribution of miles driven on four-year leases follows the normal distribution. The mean is 52,000 miles, and the standard deviation is 5,000 miles. What percent of the leases will yield a penalty because of excess mileage? (See Exercise 49 and **LO 7-3**).

## LEARNING OBJECTIVES

---

*When you have completed this chapter, you will be able to:*

- LO7-1** Describe the uniform probability distribution and use it to calculate probabilities.
- LO7-2** Describe the characteristics of a normal probability distribution.
- LO7-3** Describe the standard normal probability distribution and use it to calculate probabilities.

## INTRODUCTION

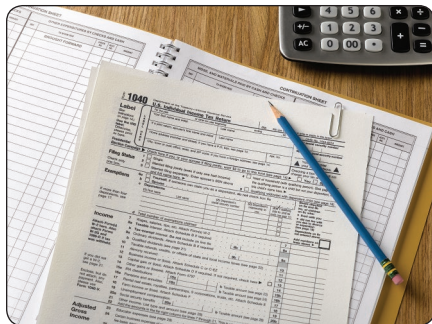
Chapter 6 began our study of probability distributions. We considered two *discrete* probability distributions: binomial and Poisson. These distributions are based on discrete random variables, which can assume only clearly separated values. For example, we select for study 10 small businesses that began operations during the year 2014. The number still operating in 2017 can be 0, 1, 2, . . . , 10. There cannot be 3.7, 12, or  $-7$  still operating in 2017. In this example, only certain outcomes are possible and these outcomes are represented by clearly separated values. In addition, the result is usually found by counting the number of successes. We count the number of businesses in the study that are still in operation in 2017.

We continue our study of probability distributions by examining *continuous* probability distributions. A continuous probability distribution usually results from measuring something, such as the distance from the dormitory to the classroom, the weight of an individual, or the amount of bonus earned by CEOs. As an example, at Dave's Inlet Fish Shack, flounder is the featured, fresh-fish menu item. The distribution of the amount of flounder sold per day has a mean of 10.0 pounds per day and a standard deviation of 3.0 pounds per day. This distribution is continuous because Dave, the owner, "measures" the amount of flounder sold each day. It is important to realize that a continuous random variable has an infinite number of values within a particular range. So, for a continuous random variable, probability is for a range of values. The probability for a specific value of a continuous random variable is 0.

This chapter shows how to use two continuous probability distributions: the uniform probability distribution and the normal probability distribution.

### LO7-1

Describe the uniform probability distribution and use it to calculate probabilities.



©Jeffrey Hamilton/Digital Vision/Getty Images RF

## THE FAMILY OF UNIFORM PROBABILITY DISTRIBUTIONS

The uniform probability distribution is the simplest distribution for a continuous random variable. This distribution is rectangular in shape and is completely defined by its minimum and maximum values. Here are some examples that follow a uniform distribution.

- The sales of gasoline at the Kwik Fill in Medina, New York, follow a uniform distribution that varies between 2,000 and 5,000 gallons per day. The random variable is the number of gallons sold per day and is continuous within the interval between 2,000 gallons and 5,000 gallons.
- Volunteers at the Grand Strand Public Library prepare federal income tax forms. The time to prepare form 1040-EZ follows a uniform distribution over the interval between 10 minutes and 30 minutes. The random variable is the number of minutes to complete the form, and it can assume any value between 10 and 30.

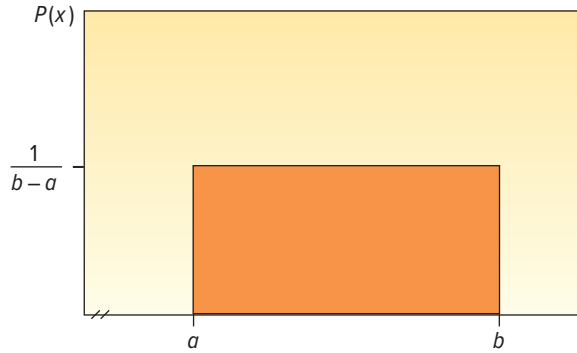
A uniform distribution is shown in Chart 7–1. The distribution's shape is rectangular and has a minimum value of  $a$  and a maximum of  $b$ . Also notice in Chart 7–1 the height of the distribution is constant or uniform for all values between  $a$  and  $b$ .

The mean of a uniform distribution is located in the middle of the interval between the minimum and maximum values. It is computed as:

**MEAN OF THE UNIFORM DISTRIBUTION**

$$\mu = \frac{a + b}{2}$$

**(7-1)**



**CHART 7-1** A Continuous Uniform Distribution

The standard deviation describes the dispersion of a distribution. In the uniform distribution, the standard deviation is also related to the interval between the maximum and minimum values.

**STANDARD DEVIATION OF THE UNIFORM DISTRIBUTION**

$$\sigma = \sqrt{\frac{(b-a)^2}{12}} \quad (7-2)$$

The equation for the uniform probability distribution is:

**UNIFORM DISTRIBUTION**  $P(x) = \frac{1}{b-a}$  if  $a \leq x \leq b$  and 0 elsewhere **(7-3)**

As we described in Chapter 6, probability distributions are useful for making probability statements concerning the values of a random variable. For distributions describing a continuous random variable, areas within the distribution represent probabilities. In the uniform distribution, its rectangular shape allows us to apply the area formula for a rectangle. Recall that we find the area of a rectangle by multiplying its length by its height. For the uniform distribution, the height of the rectangle is  $P(x)$ , which is  $1/(b-a)$ . The length or base of the distribution is  $b-a$ . So if we multiply the height of the distribution by its entire range to find the area, the result is always 1.00. To put it another way, the total area within a continuous probability distribution is equal to 1.00. In general

$$\text{Area} = (\text{height})(\text{base}) = \frac{1}{(b-a)} (b-a) = 1.00$$

So if a uniform distribution ranges from 10 to 15, the height is 0.20, found by  $1/(15-10)$ . The base is 5, found by  $15-10$ . The total area is:

$$\text{Area} = (\text{height})(\text{base}) = \frac{1}{(15-10)} (15-10) = 1.00$$

The following example illustrates the features of a uniform distribution and how we use it to calculate probabilities.

**EXAMPLE**

Southwest Arizona State University provides bus service to students while they are on campus. A bus arrives at the North Main Street and College Drive stop every 30 minutes between 6 a.m. and 11 p.m. during weekdays. Students arrive at the

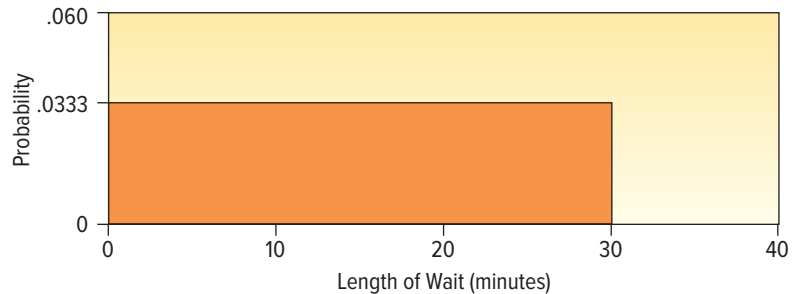
bus stop at random times. The time that a student waits is uniformly distributed from 0 to 30 minutes.

1. Draw a graph of this distribution.
2. Show that the area of this uniform distribution is 1.00.
3. How long will a student “typically” have to wait for a bus? In other words, what is the mean waiting time? What is the standard deviation of the waiting times?
4. What is the probability a student will wait more than 25 minutes?
5. What is the probability a student will wait between 10 and 20 minutes?

### SOLUTION

In this case, the random variable is the length of time a student must wait. Time is measured on a continuous scale, and the wait times may range from 0 minutes up to 30 minutes.

1. The graph of the uniform distribution is shown in Chart 7–2. The horizontal line is drawn at a height of .0333, found by  $1/(30 - 0)$ . The range of this distribution is 30 minutes.



**CHART 7–2** Uniform Probability Distribution of Student Waiting Times

2. The times students must wait for the bus are uniform over the interval from 0 minutes to 30 minutes, so in this case  $a$  is 0 and  $b$  is 30.

$$\text{Area} = (\text{height})(\text{base}) = \frac{1}{(30 - 0)} (30 - 0) = 1.00$$

3. To find the mean, we use formula (7–1).

$$\mu = \frac{a + b}{2} = \frac{0 + 30}{2} = 15$$

The mean of the distribution is 15 minutes, so the typical wait time for bus service is 15 minutes.

To find the standard deviation of the wait times, we use formula (7–2).

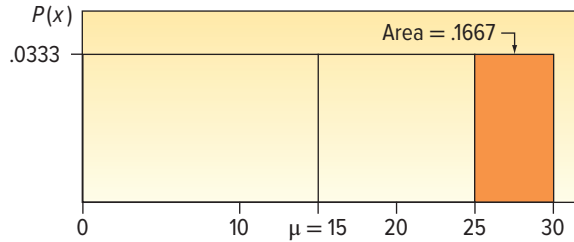
$$\sigma = \sqrt{\frac{(b - a)^2}{12}} = \sqrt{\frac{(30 - 0)^2}{12}} = 8.66$$

The standard deviation of the distribution is 8.66 minutes. This measures the variation in the student wait times.

4. The area within the distribution for the interval 25 to 30 represents this particular probability. From the area formula:

$$P(25 < \text{wait time} < 30) = (\text{height})(\text{base}) = \frac{1}{(30 - 0)} (5) = .1667$$

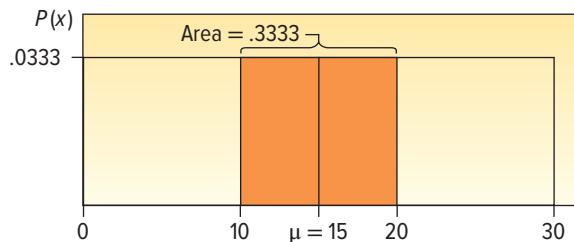
So the probability a student waits between 25 and 30 minutes is .1667. This conclusion is illustrated by the following graph.



5. The area within the distribution for the interval 10 to 20 represents the probability.

$$P(10 < \text{wait time} < 20) = (\text{height})(\text{base}) = \frac{1}{(30 - 0)} (10) = .3333$$

We can illustrate this probability as follows.



## SELF-REVIEW 7-1



Microwave ovens only last so long. The useful life of a microwave oven follows a uniform distribution between 8 and 14 years.

- Draw this uniform distribution. What are the height and base values?
- Show the total area under the curve is 1.00.
- Calculate the mean and the standard deviation of this distribution.
- What is the probability a particular microwave oven lasts between 10 and 14 years?
- What is the probability a microwave oven will last less than 9 years?

## EXERCISES

- A uniform distribution is defined over the interval from 6 to 10.
  - What are the values for  $a$  and  $b$ ?
  - What is the mean of this uniform distribution?
  - What is the standard deviation?
  - Show that the total area is 1.00.
  - Find the probability of a value more than 7.
  - Find the probability of a value between 7 and 9.
- A uniform distribution is defined over the interval from 2 to 5.
  - What are the values for  $a$  and  $b$ ?
  - What is the mean of this uniform distribution?
  - What is the standard deviation?
  - Show that the total area is 1.00.
  - Find the probability of a value more than 2.6.
  - Find the probability of a value between 2.9 and 3.7.

3. The closing price of Schnur Sporting Goods Inc. common stock is uniformly distributed between \$20 and \$30 per share. What is the probability that the stock price will be:
  - a. More than \$27?
  - b. Less than \$24?
4. According to the Insurance Institute of America, a family of four spends between \$400 and \$3,800 per year on all types of insurance. Suppose the money spent is uniformly distributed between these amounts.
  - a. What is the mean amount spent on insurance?
  - b. What is the standard deviation of the amount spent?
  - c. If we select a family at random, what is the probability they spend less than \$2,000 per year on insurance?
  - d. What is the probability a family spends more than \$3,000 per year?
5. The April rainfall in Flagstaff, Arizona, follows a uniform distribution between 0.5 and 3.00 inches.
  - a. What are the values for  $a$  and  $b$ ?
  - b. What is the mean amount of rainfall for the month? What is the standard deviation?
  - c. What is the probability of less than an inch of rain for the month?
  - d. What is the probability of *exactly* 1.00 inch of rain?
  - e. What is the probability of more than 1.50 inches of rain for the month?
6. Customers experiencing technical difficulty with their Internet cable service may call an 800 number for technical support. It takes the technician between 30 seconds and 10 minutes to resolve the problem. The distribution of this support time follows the uniform distribution.
  - a. What are the values for  $a$  and  $b$  in minutes?
  - b. What is the mean time to resolve the problem? What is the standard deviation of the time?
  - c. What percent of the problems take more than 5 minutes to resolve?
  - d. Suppose we wish to find the middle 50% of the problem-solving times. What are the end points of these two times?

**LO7-2**

Describe the characteristics of a normal probability distribution.

## THE FAMILY OF NORMAL PROBABILITY DISTRIBUTIONS

Next we consider the normal probability distribution. Unlike the uniform distribution [see formula (7–3)], the normal probability distribution has a very complex formula.

### NORMAL PROBABILITY DISTRIBUTION

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]} \quad (7-4)$$

However, do not be bothered by how complex this formula looks. You are already familiar with many of the values. The symbols  $\mu$  and  $\sigma$  refer to the mean and the standard deviation, as usual. The Greek symbol  $\pi$  is a constant and its value is approximately 22/7 or 3.1416. The letter  $e$  is also a constant. It is the base of the natural log system and is approximately equal to 2.718.  $x$  is the value of a continuous random variable. So a normal distribution is based on—that is, it is defined by—its mean and standard deviation.

You will not need to make calculations using formula (7–4). Instead you will use a table, given in Appendix B.3, to find various probabilities. These probabilities can also be calculated using Excel functions as well as other statistical software.

## STATISTICS IN ACTION

Many variables are approximately, normally distributed, such as IQ scores, life expectancies, and adult height. This implies that nearly all observations occur within 3 standard deviations of the mean. On the other hand, observations that occur beyond 3 standard deviations from the mean are extremely rare. For example, the mean adult male height is 68.2 inches (about 5 feet 8 inches) with a standard deviation of 2.74. This means that almost all males are between 60.0 inches (5 feet) and 76.4 inches (6 feet 4 inches). LeBron James, a professional basketball player with the Cleveland Cavaliers, is 80 inches, or 6 feet 8 inches, which is clearly beyond 3 standard deviations from the mean. The height of a standard doorway is 6 feet 8 inches, which should be high enough for almost all adult males, except for a rare person like LeBron James.

As another example, the driver's seat in most vehicles is set to comfortably fit a person who is at least 159 cm (62.5 inches) tall. The distribution of heights of adult women is approximately a normal distribution with a mean of 161.5 cm and a standard deviation of 6.3 cm. Thus about 35% of adult women will not fit comfortably in the driver's seat.

The normal probability distribution has the following characteristics:

- It is **bell-shaped** and has a single peak at the center of the distribution. The arithmetic mean, median, and mode are equal and located in the center of the distribution. The total area under the curve is 1.00. Half the area under the normal curve is to the right of this center point, and the other half is to the left of it.
- It is **symmetrical** about the mean. If we cut the normal curve vertically at the center value, the shapes of the curves will be mirror images. Also, the area of each half is 0.5.
- It falls off smoothly in either direction from the central value. That is, the distribution is **asymptotic**: The curve gets closer and closer to the X-axis but never actually touches it. To put it another way, the tails of the curve extend indefinitely in both directions.
- The location of a normal distribution is determined by the mean,  $\mu$ . The dispersion or spread of the distribution is determined by the standard deviation,  $\sigma$ .

These characteristics are shown graphically in Chart 7–3.

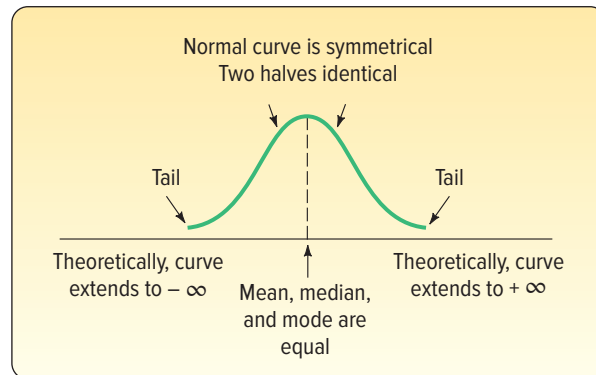


CHART 7–3 Characteristics of a Normal Distribution

There is not just one normal probability distribution, but rather a “family” of them. For example, in Chart 7–4 the probability distributions of length of employee service in three different plants are compared. In the Camden plant, the mean is 20 years and the standard deviation is 3.1 years. There is another normal probability distribution for the length of service in the Dunkirk plant, where  $\mu = 20$  years and  $\sigma = 3.9$  years. In the Elmira plant,  $\mu = 20$  years and  $\sigma = 5.0$  years. Note that the means are the same but the standard deviations are different. As the standard deviation gets smaller, the distribution becomes more narrow and “peaked.”

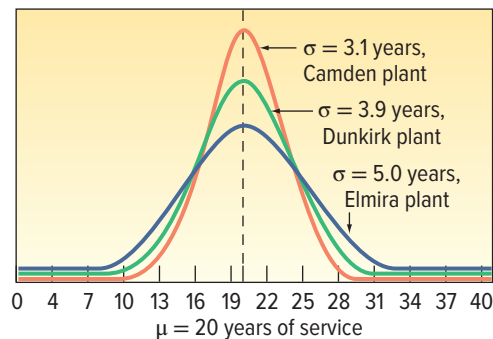
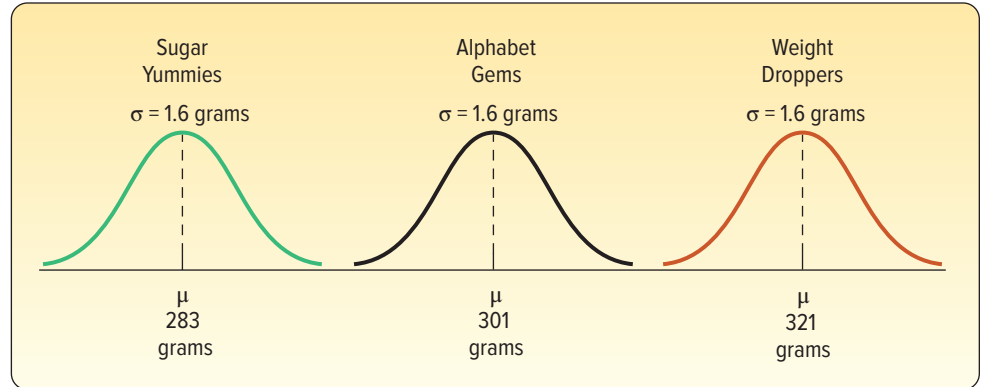


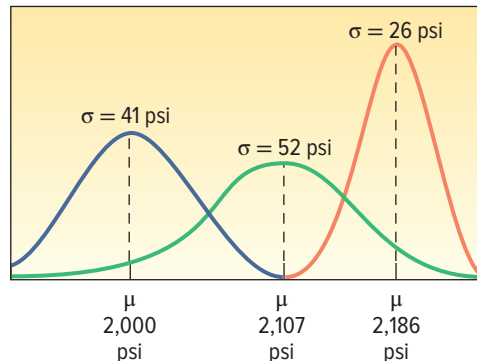
CHART 7–4 Normal Probability Distributions with Equal Means but Different Standard Deviations



**CHART 7-5** Normal Probability Distributions Having Different Means but Equal Standard Deviations

Chart 7-5 shows the distribution of box weights of three different cereals. The weights follow a normal distribution with different means but identical standard deviations.

Finally, Chart 7-6 shows three normal distributions having different means and standard deviations. They show the distribution of tensile strengths, measured in pounds per square inch (psi), for three types of cables.



**CHART 7-6** Normal Probability Distributions with Different Means and Standard Deviations

In Chapter 6, recall that discrete probability distributions show the specific likelihood a discrete value will occur. For example, on page 166, the binomial distribution is used to calculate the probability that none of the five flights arriving at Pennsylvania's Bradford Regional Airport will be late.

With a continuous probability distribution, areas below the curve define probabilities. The total area under the normal curve is 1.0. This accounts for all possible outcomes. Because a normal probability distribution is symmetric, the area under the curve to the left of the mean is 0.5, and the area under the curve to the right of the mean is 0.5. Apply this to the distribution of Sugar Yummies in Chart 7-5. It is normally distributed with a mean of 283 grams. Therefore, the probability of filling a box with more than 283 grams is 0.5 and the probability of filling a box with less than 283 grams is 0.5. We also can determine the probability that a box weighs between 280 and 286 grams. However, to determine this probability we need to know about the standard normal probability distribution.



**LO7-3**

Describe the standard normal probability distribution and use it to calculate probabilities.

## THE STANDARD NORMAL PROBABILITY DISTRIBUTION

The number of normal distributions is unlimited, each having a different mean ( $\mu$ ), standard deviation ( $\sigma$ ), or both. While it is possible to provide a limited number of probability tables for discrete distributions such as the binomial and the Poisson, providing tables for the infinite number of normal distributions is impractical. Fortunately, one member of the family can be used to determine the probabilities for all normal probability distributions. It is called the **standard normal probability distribution**, and it is unique because it has a mean of 0 and a standard deviation of 1.

Any *normal probability distribution* can be converted into a *standard normal probability distribution* by subtracting the mean from each observation and dividing this difference by the standard deviation. The results are called **z values** or **z scores**.

**z VALUE** The signed distance between a selected value, designated  $x$ , and the mean,  $\mu$ , divided by the standard deviation,  $\sigma$ .

So, a z value is the distance from the mean, measured in units of the standard deviation. The formula for this conversion is:

**STANDARD NORMAL VALUE**

$$z = \frac{x - \mu}{\sigma} \quad (7-5)$$

where:

- $x$  is the value of any particular observation or measurement.
- $\mu$  is the mean of the distribution.
- $\sigma$  is the standard deviation of the distribution.

As we noted in the preceding definition, a z value expresses the distance or difference between a particular value of  $x$  and the arithmetic mean in units of the standard deviation. Once the normally distributed observations are standardized, the z values are normally distributed with a mean of 0 and a standard deviation of 1. Therefore, the z distribution has all the characteristics of any normal probability distribution. These characteristics are listed on page 190 in the Family of Normal Probability Distributions section. The table in Appendix B.3 lists the probabilities for the standard normal probability distribution. Table 7–1 presents a small portion of this table.

**TABLE 7–1** Areas under the Normal Curve

| z   | 0.00   | 0.01   | 0.02   | 0.03   | 0.04   | 0.05   | ... |
|-----|--------|--------|--------|--------|--------|--------|-----|
| 1.3 | 0.4032 | 0.4049 | 0.4066 | 0.4082 | 0.4099 | 0.4115 |     |
| 1.4 | 0.4192 | 0.4207 | 0.4222 | 0.4236 | 0.4251 | 0.4265 |     |
| 1.5 | 0.4332 | 0.4345 | 0.4357 | 0.4370 | 0.4382 | 0.4394 |     |
| 1.6 | 0.4452 | 0.4463 | 0.4474 | 0.4484 | 0.4495 | 0.4505 |     |
| 1.7 | 0.4554 | 0.4564 | 0.4573 | 0.4582 | 0.4591 | 0.4599 |     |
| 1.8 | 0.4641 | 0.4649 | 0.4656 | 0.4664 | 0.4671 | 0.4678 |     |
| 1.9 | 0.4713 | 0.4719 | 0.4726 | 0.4732 | 0.4738 | 0.4744 |     |
| .   |        |        |        |        |        |        |     |
| :   |        |        |        |        |        |        |     |
| :   |        |        |        |        |        |        |     |

### STATISTICS IN ACTION

An individual's skills depend on a combination of many hereditary and environmental factors, each having about the same amount of weight or influence on the skills. Thus, much like a binomial distribution with a large number of trials, many skills and attributes follow the normal distribution. For example, the SAT Reasoning Test is the most widely used standardized test for college admissions in the United States. Scores are based on a normal distribution with a mean of 1,500 and a standard deviation of 300.

## Applications of the Standard Normal Distribution

The standard normal distribution is very useful for determining probabilities for any normally distributed random variable. The basic procedure is to find the  $z$  value for a particular value of the random variable based on the mean and standard deviation of its distribution. Then, using the  $z$  value, we can use the standard normal distribution to find various probabilities. The following example/solution describes the details of the application.

### EXAMPLE

In recent years a new type of taxi service has evolved in more than 300 cities worldwide, where the customer is connected directly with a driver via a smartphone. The idea was first developed by Uber Technologies, which is headquartered in San Francisco, California. It uses the Uber mobile app, which allows customers with a smartphone to submit a trip request, which is then routed to an Uber driver who picks up the customer and takes the customer to the desired location. No cash is involved; the payment for the transaction is handled via a digital payment.

Suppose the weekly income of Uber drivers follows the normal probability distribution with a mean of \$1,000 and a standard deviation of \$100. What is the  $z$  value of income for a driver who earns \$1,100 per week? For a driver who earns \$900 per week?

### SOLUTION

Using formula (7–5), the  $z$  values corresponding to the two  $x$  values (\$1,100 and \$900) are:

$$\begin{array}{ll} \text{For } x = \$1,100: & \text{For } x = \$900: \\ z = \frac{x - \mu}{\sigma} & z = \frac{x - \mu}{\sigma} \\ = \frac{\$1,100 - \$1,000}{\$100} & = \frac{\$900 - \$1,000}{\$100} \\ = 1.00 & = -1.00 \end{array}$$

The  $z$  of 1.00 indicates that a weekly income of \$1,100 is one standard deviation above the mean, and a  $z$  of  $-1.00$  shows that a \$900 income is one standard deviation below the mean. Note that both incomes (\$1,100 and \$900) are the same distance (\$100) from the mean.

## SELF-REVIEW 7-2



A recent national survey concluded that the typical person consumes 48 ounces of water per day. Assume daily water consumption follows a normal probability distribution with a standard deviation of 12.8 ounces.

- What is the  $z$  value for a person who consumes 64 ounces of water per day? Based on this  $z$  value, how does this person compare to the national average?
- What is the  $z$  value for a person who consumes 32 ounces of water per day? Based on this  $z$  value, how does this person compare to the national average?

## The Empirical Rule

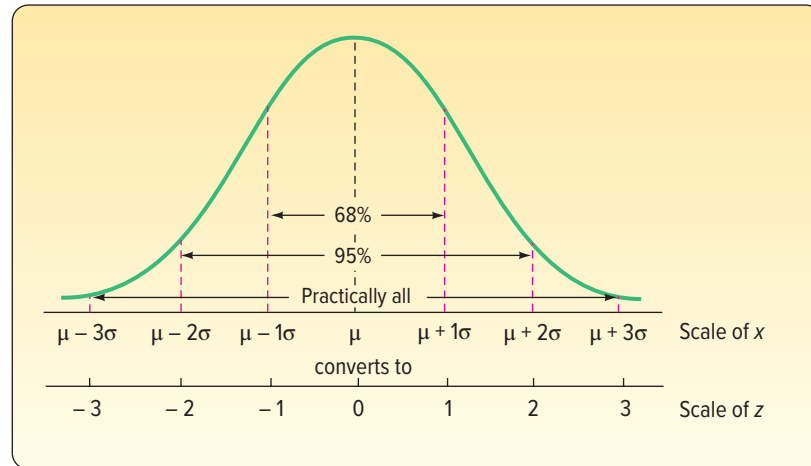
The Empirical Rule is introduced on page 79 of Chapter 3. It states that if a random variable is normally distributed, then:

- Approximately 68% of the observations will lie within plus and minus one standard deviation of the mean.

2. About 95% of the observations will lie within plus and minus two standard deviations of the mean.
3. Practically all, or 99.7% of the observations, will lie within plus and minus three standard deviations of the mean.

Now, knowing how to apply the standard normal probability distribution, we can verify the Empirical Rule. For example, one standard deviation from the mean is the same as a  $z$  value of 1.00. When we refer to the standard normal probability table, a  $z$  value of 1.00 corresponds to a probability of 0.3413. So what percent of the observations will lie within plus and minus one standard deviation of the mean? We multiply  $(2)(0.3413)$ , which equals 0.6826, or approximately 68% of the observations are within plus and minus one standard deviation of the mean.

The Empirical Rule is summarized in the following graph.



Transforming measurements to standard normal deviates changes the scale. The conversions are also shown in the graph. For example,  $\mu + 1\sigma$  is converted to a  $z$  value of 1.00. Likewise,  $\mu - 2\sigma$  is transformed to a  $z$  value of  $-2.00$ . Note that the center of the  $z$  distribution is zero, indicating no deviation from the mean,  $\mu$ .

### EXAMPLE

As part of its quality assurance program, the Autolite Battery Company conducts tests on battery life. For a particular D-cell alkaline battery, the mean life is 19 hours. The useful life of the battery follows a normal distribution with a standard deviation of 1.2 hours. Answer the following questions.

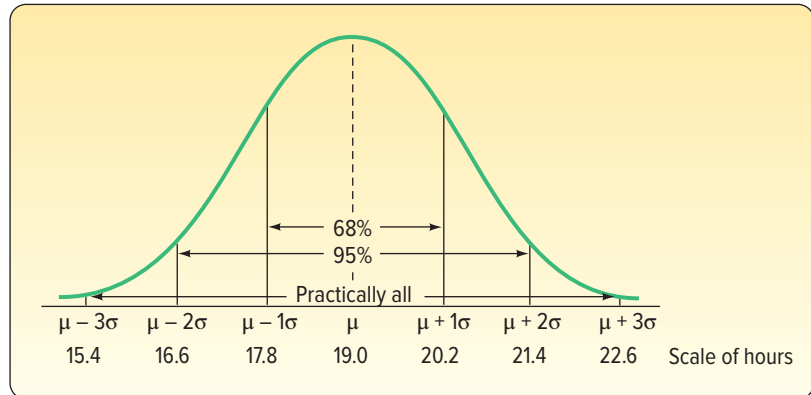
1. About 68% of batteries have a life between what two values?
2. About 95% of batteries have a life between what two values?
3. Virtually all, or 99%, of batteries have a life between what two values?

### SOLUTION

We can use the Empirical Rule to answer these questions.

1. We can expect about 68% of the batteries to last between 17.8 and 20.2 hours, found by  $19.0 \pm 1(1.2)$  hours.
2. We can expect about 95% of the batteries to last between 16.6 and 21.4 hours, found by  $19.0 \pm 2(1.2)$  hours.
3. We can expect about 99%, or practically all, of the batteries to last between 15.4 and 22.6 hours, found by  $19.0 \pm 3(1.2)$  hours.

This information is summarized on the following chart.



## SELF-REVIEW 7-3



The distribution of the annual incomes of a group of middle-management employees at Compton Plastics approximates a normal distribution with a mean of \$47,200 and a standard deviation of \$800.

- About 68% of the incomes lie between what two amounts?
- About 95% of the incomes lie between what two amounts?
- Virtually all of the incomes lie between what two amounts?
- What are the median and the modal incomes?
- Is the distribution of incomes symmetrical?

## EXERCISES

- Explain what is meant by this statement: "There is not just one normal probability distribution but a 'family' of them."
- List the major characteristics of a normal probability distribution.
- The mean of a normal probability distribution is 500; the standard deviation is 10.
  - About 68% of the observations lie between what two values?
  - About 95% of the observations lie between what two values?
  - Practically all of the observations lie between what two values?
- The mean of a normal probability distribution is 60; the standard deviation is 5.
  - About what percent of the observations lie between 55 and 65?
  - About what percent of the observations lie between 50 and 70?
  - About what percent of the observations lie between 45 and 75?
- The Kamp family has twins, Rob and Rachel. Both Rob and Rachel graduated from college 2 years ago, and each is now earning \$50,000 per year. Rachel works in the retail industry, where the mean salary for executives with less than 5 years' experience is \$35,000 with a standard deviation of \$8,000. Rob is an engineer. The mean salary for engineers with less than 5 years' experience is \$60,000 with a standard deviation of \$5,000. Compute the  $z$  values for both Rob and Rachel, and comment on your findings.
- A recent article in the *Cincinnati Enquirer* reported that the mean labor cost to repair a heat pump is \$90 with a standard deviation of \$22. Monte's Plumbing and Heating Service completed repairs on two heat pumps this morning. The labor cost for the first was \$75, and it was \$100 for the second. Assume the distribution of labor costs follows the normal probability distribution. Compute  $z$  values for each, and comment on your findings.

## Finding Areas under the Normal Curve

The next application of the standard normal distribution involves finding the area in a normal distribution between the mean and a selected value, which we identify as  $x$ . The following example/solution will illustrate the details.

### EXAMPLE

In the first example/solution described on page 193 in this section, we reported that the weekly income of Uber drivers followed the normal distribution with a mean of \$1,000 and a standard deviation of \$100. That is,  $\mu = \$1,000$  and  $\sigma = \$100$ . What is the likelihood of selecting a driver whose weekly income is between \$1,000 and \$1,100?

### SOLUTION

We have already converted \$1,100 to a  $z$  value of 1.00 using formula (7-5). To repeat:

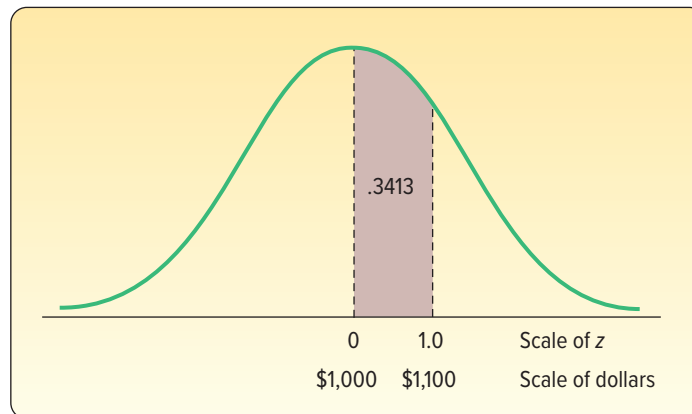
$$z = \frac{x - \mu}{\sigma} = \frac{\$1,100 - \$1,000}{\$100} = 1.00$$

The probability associated with a  $z$  of 1.00 is available in Appendix B.3. A portion of Appendix B.3 follows. To locate the probability, go down the left column to 1.0, and then move horizontally to the column headed 0.00. The value is .3413.

| $z$ | 0.00  | 0.01  | 0.02  |
|-----|-------|-------|-------|
| ⋮   | ⋮     | ⋮     | ⋮     |
| ⋮   | ⋮     | ⋮     | ⋮     |
| 0.7 | .2580 | .2611 | .2642 |
| 0.8 | .2881 | .2910 | .2939 |
| 0.9 | .3159 | .3186 | .3212 |
| 1.0 | .3413 | .3438 | .3461 |
| 1.1 | .3643 | .3665 | .3686 |
| ⋮   | ⋮     | ⋮     | ⋮     |
| ⋮   | ⋮     | ⋮     | ⋮     |

The area under the normal curve between \$1,000 and \$1,100 is .3413. We could also say 34.13% of Uber drivers earn between \$1,000 and \$1,100 weekly, or the likelihood of selecting a driver and finding his or her income is between \$1,000 and \$1,100 is .3413.

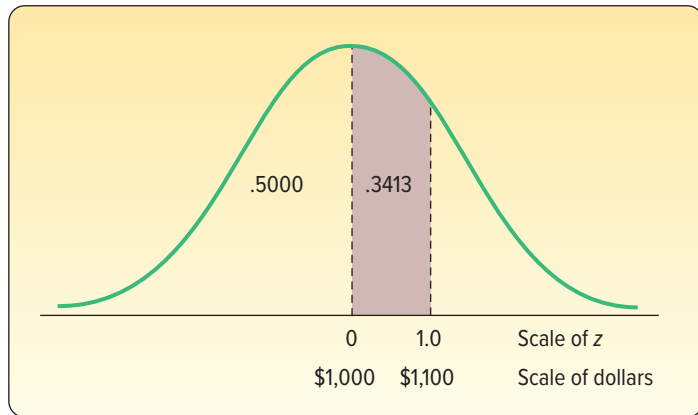
This information is summarized in the following diagram.



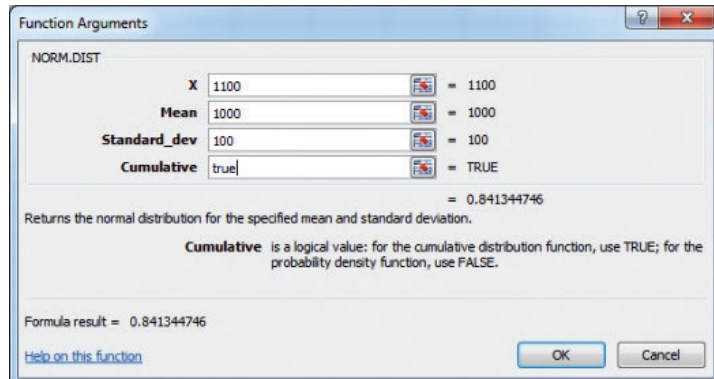
**STATISTICS IN ACTION**

Many processes, such as filling soda bottles and canning fruit, are normally distributed. Manufacturers must guard against both over- and underfilling. If they put too much in the can or bottle, they are giving away their product. If they put too little in, the customer may feel cheated and the government may question the label description. “Control charts,” with limits drawn three standard deviations above and below the mean, are routinely used to monitor this type of production process.

In the example/solution just completed, we are interested in the probability between the mean and a given value. Let’s change the question. Instead of wanting to know the probability of selecting a random driver who earned between \$1,000 and \$1,100, suppose we wanted the probability of selecting a driver who earned less than \$1,100. In probability notation, we write this statement as  $P(\text{weekly income} < \$1,100)$ . The method of solution is the same. We find the probability of selecting a driver who earns between \$1,000, the mean, and \$1,100. This probability is .3413. Next, recall that half the area, or probability, is above the mean and half is below. So the probability of selecting a driver earning less than \$1,000 is .5000. Finally, we add the two probabilities, so  $.3413 + .5000 = .8413$ . About 84% of Uber drivers earn less than \$1,100 per week. See the following diagram.



Excel will calculate this probability. The necessary commands are in the **Software Commands** in Appendix C. The answer is .8413, the same as we calculated.



Source: Microsoft Excel

**EXAMPLE**

Refer to the first example/solution discussed on page 193 in this section regarding the weekly income of Uber drivers. The distribution of weekly incomes follows the normal probability distribution, with a mean of \$1,000 and a standard deviation of \$100. What is the probability of selecting a driver whose income is:

1. Between \$790 and \$1,000?
2. Less than \$790?

**SOLUTION**

We begin by finding the  $z$  value corresponding to a weekly income of \$790. From formula (7-5):

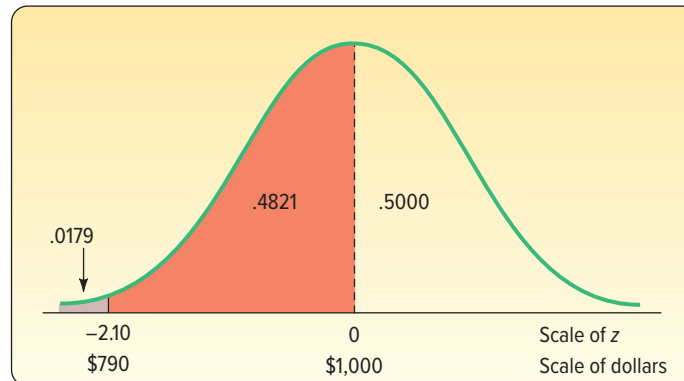
$$z = \frac{x - \mu}{s} = \frac{\$790 - \$1,000}{\$100} = -2.10$$

See Appendix B.3. Move down the left margin to row 2.1 and across that row to the column headed 0.00. The value is .4821. So the area under the standard normal curve corresponding to a  $z$  value of 2.10 is .4821. However, because the normal distribution is symmetric, the area between 0 and a negative  $z$  value is the same as that between 0 and the corresponding positive  $z$  value. The likelihood of finding a driver earning between \$790 and \$1,000 is .4821. In probability notation, we write  $P(\$790 < \text{weekly income} < \$1,000) = .4821$ .

| $z$ | 0.00  | 0.01  | 0.02  |
|-----|-------|-------|-------|
| .   | .     | .     | .     |
| .   | .     | .     | .     |
| .   | .     | .     | .     |
| 2.0 | .4772 | .4778 | .4783 |
| 2.1 | .4821 | .4826 | .4830 |
| 2.2 | .4861 | .4864 | .4868 |
| 2.3 | .4893 | .4896 | .4898 |
| .   | .     | .     | .     |
| .   | .     | .     | .     |

The mean divides the normal curve into two identical halves. The area under the half to the left of the mean is .5000, and the area to the right is also .5000. Because the area under the curve between \$790 and \$1,000 is .4821, the area below \$790 is .0179, found by  $.5000 - .4821$ . In probability notation, we write  $P(\text{weekly income} < \$790) = .0179$ .

So we conclude that 48.21% of Uber drivers have weekly incomes between \$790 and \$1,000. Further, we can anticipate that 1.79% earn less than \$790 per week. This information is summarized in the following diagram.

**SELF-REVIEW 7-4**

The temperature of coffee sold at the Coffee Bean Cafe follows the normal probability distribution, with a mean of 150 degrees. The standard deviation of this distribution is 5 degrees.

- What is the probability that the coffee temperature is between 150 degrees and 154 degrees?
- What is the probability that the coffee temperature is more than 164 degrees?

## EXERCISES

13. A normal population has a mean of 20.0 and a standard deviation of 4.0.
  - a. Compute the  $z$  value associated with 25.0.
  - b. What proportion of the population is between 20.0 and 25.0?
  - c. What proportion of the population is less than 18.0?
14. A normal population has a mean of 12.2 and a standard deviation of 2.5.
  - a. Compute the  $z$  value associated with 14.3.
  - b. What proportion of the population is between 12.2 and 14.3?
  - c. What proportion of the population is less than 10.0?
15. A recent study of the hourly wages of maintenance crew members for major airlines showed that the mean hourly wage was \$20.50, with a standard deviation of \$3.50. Assume the distribution of hourly wages follows the normal probability distribution. If we select a crew member at random, what is the probability the crew member earns:
  - a. Between \$20.50 and \$24.00 per hour?
  - b. More than \$24.00 per hour?
  - c. Less than \$19.00 per hour?
16. The mean of a normal probability distribution is 400 pounds. The standard deviation is 10 pounds.
  - a. What is the area between 415 pounds and the mean of 400 pounds?
  - b. What is the area between the mean and 395 pounds?
  - c. What is the probability of selecting a value at random and discovering that it has a value of less than 395 pounds?

Another application of the normal distribution involves combining two areas, or probabilities. One of the areas is to the right of the mean and the other to the left.

### EXAMPLE

Continuing the example/solution first discussed on page 193 using the weekly income of Uber drivers, weekly income follows the normal probability distribution, with a mean of \$1,000 and a standard deviation of \$100. What is the area under this normal curve between \$840 and \$1,200?

### SOLUTION

The problem can be divided into two parts. For the area between \$840 and the mean of \$1,000:

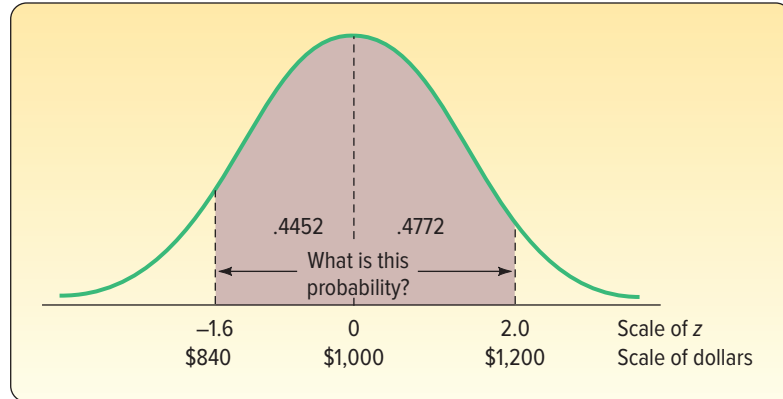
$$z = \frac{\$840 - \$1,000}{\$100} = \frac{-\$160}{\$100} = -1.60$$

For the area between the mean of \$1,000 and \$1,200:

$$z = \frac{\$1,200 - \$1,000}{\$100} = \frac{\$200}{\$100} = 2.00$$

The area under the curve for a  $z$  of  $-1.60$  is .4452 (from Appendix B.3). The area under the curve for a  $z$  of 2.00 is .4772. Adding the two areas:  $.4452 + .4772 = .9224$ . Thus, the probability of selecting an income between \$840 and \$1,200 is .9224. In probability notation, we write  $P(\$840 < \text{weekly income} < \$1,200) = .4452 + .4772 = .9224$ . To summarize, 92.24% of the drivers have weekly incomes between \$840 and \$1,200. This is shown in a diagram:





Another application of the normal distribution involves determining the area between values on the *same* side of the mean.

### EXAMPLE

Returning to the weekly income distribution of Uber drivers ( $\mu = \$1,000$ ,  $\sigma = \$100$ ), what is the area under the normal curve between \$1,150 and \$1,250?

### SOLUTION

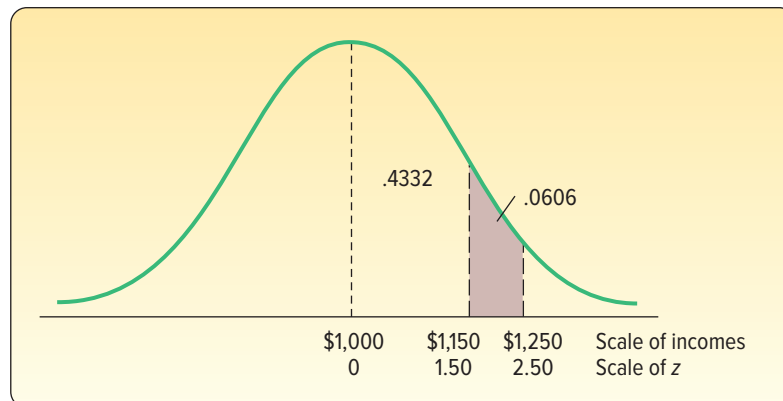
The situation is again separated into two parts, and formula (7-5) is used. First, we find the z value associated with a weekly income of \$1,250:

$$z = \frac{\$1,250 - \$1,000}{\$100} = 2.50$$

Next we find the z value for a weekly income of \$1,150:

$$z = \frac{\$1,150 - \$1,000}{\$100} = 1.50$$

From Appendix B.3, the area associated with a z value of 2.50 is .4938. So the probability of a weekly income between \$1,000 and \$1,250 is .4938. Similarly, the area associated with a z value of 1.50 is .4332, so the probability of a weekly income between \$1,000 and \$1,150 is .4332. The probability of a weekly income between \$1,150 and \$1,250 is found by subtracting the area associated with a z value of 1.50 (.4332) from that associated with a z of 2.50 (.4938). Thus, the probability of a weekly income between \$1,150 and \$1,250 is .0606. In probability notation, we write  $P(\$1,150 < \text{weekly income} < \$1,250) = .4938 - .4332 = .0606$ .



To summarize, there are four situations for finding the area under the standard normal probability distribution.

1. To find the area between 0 and  $z$  or  $(-z)$ , look up the probability directly in the table.
2. To find the area beyond  $z$  or  $(-z)$ , locate the probability of  $z$  in the table and subtract that probability from .5000.
3. To find the area between two points on different sides of the mean, determine the  $z$  values and add the corresponding probabilities.
4. To find the area between two points on the same side of the mean, determine the  $z$  values and subtract the smaller probability from the larger.

## SELF-REVIEW 7-5



Refer to Self-Review 7-4. The temperature of coffee sold at the Coffee Bean Cafe follows the normal probability distribution with a mean of 150 degrees. The standard deviation of this distribution is 5 degrees.

- (a) What is the probability the coffee temperature is between 146 degrees and 156 degrees?
- (b) What is the probability the coffee temperature is more than 156 but less than 162 degrees?

## EXERCISES

17. A normal distribution has a mean of 50 and a standard deviation of 4.
  - a. Compute the probability of a value between 44.0 and 55.0.
  - b. Compute the probability of a value greater than 55.0.
  - c. Compute the probability of a value between 52.0 and 55.0.
18. A normal population has a mean of 80.0 and a standard deviation of 14.0.
  - a. Compute the probability of a value between 75.0 and 90.0.
  - b. Compute the probability of a value of 75.0 or less.
  - c. Compute the probability of a value between 55.0 and 70.0.
19. Suppose the Internal Revenue Service reported that the mean tax refund for the year 2017 was \$2,800. Assume the standard deviation is \$450 and that the amounts refunded follow a normal probability distribution.
  - a. What percent of the refunds are more than \$3,100?
  - b. What percent of the refunds are more than \$3,100 but less than \$3,500?
  - c. What percent of the refunds are more than \$2,250 but less than \$3,500?
20. The distribution of the number of viewers for the *American Idol* television show follows a normal distribution with a mean of 29 million and a standard deviation of 5 million. What is the probability next week's show will:
  - a. Have between 30 and 34 million viewers?
  - b. Have at least 23 million viewers?
  - c. Exceed 40 million viewers?
21. WNAE, an all-news AM station, finds that the distribution of the lengths of time listeners are tuned to the station follows the normal distribution. The mean of the distribution is 15.0 minutes and the standard deviation is 3.5 minutes. What is the probability that a particular listener will tune in for:
  - a. More than 20 minutes?
  - b. 20 minutes or less?
  - c. Between 10 and 12 minutes?
22. Among the thirty largest U.S. cities, the mean one-way commute time to work is 25.8 minutes. The longest one-way travel time is in New York City, where the mean time is 39.7 minutes. Assume the distribution of travel times in New York City follows the normal probability distribution and the standard deviation is 7.5 minutes.
  - a. What percent of the New York City commutes are for less than 30 minutes?
  - b. What percent are between 30 and 35 minutes?
  - c. What percent are between 30 and 50 minutes?

The previous example/solutions require finding the percent of the observations located between two observations or the percent of the observations above, or below, a particular observation  $x$ . A further application of the normal distribution involves finding the value of the observation  $x$  when the percent above or below the observation is given.

### EXAMPLE

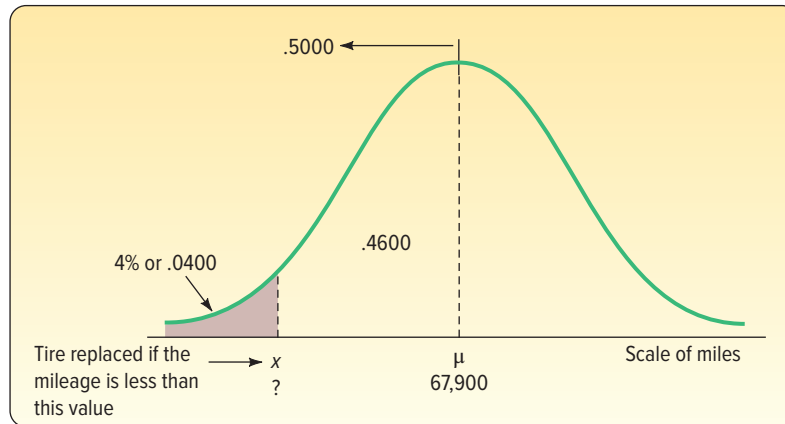
Layton Tire and Rubber Company wishes to set a minimum mileage guarantee on its new MX100 tire. Tests reveal the mean mileage is 67,900 with a standard deviation of 2,050 miles and that the distribution of miles follows the normal probability distribution. Layton wants to set the minimum guaranteed mileage so that no more than 4% of the tires will have to be replaced. What minimum guaranteed mileage should Layton announce?

### SOLUTION

The facets of this case are shown in the following diagram, where  $x$  represents the minimum guaranteed mileage.



©Jupiterimages/Getty Images RF



Inserting these values in formula (7-5) for  $z$  gives:

$$z = \frac{x - \mu}{\sigma} = \frac{x - 67,900}{2,050}$$

There are two unknowns in this equation,  $z$  and  $x$ . To find  $x$ , we first find  $z$  and then solve for  $x$ . Recall from the characteristics of a normal curve that the area to the left of  $\mu$  is .5000. The area between  $\mu$  and  $x$  is .4600, found by .5000 - .0400. Now refer to Appendix B.3. Search the body of the table for the area closest to .4600. The closest area is .4599. Move to the margins from this value and read

the z value of 1.75. Because the value is to the left of the mean, it is actually -1.75. These steps are illustrated in Table 7-2.

**TABLE 7-2** Selected Areas under the Normal Curve

| z ... | .03   | .04   | .05   | .06   |
|-------|-------|-------|-------|-------|
| ...   | ...   | ...   | ...   | ...   |
| 1.5   | .4370 | .4382 | .4394 | .4406 |
| 1.6   | .4484 | .4495 | .4505 | .4515 |
| 1.7   | .4582 | .4591 | .4599 | .4608 |
| 1.8   | .4664 | .4671 | .4678 | .4686 |

Knowing that the distance between  $\mu$  and  $x$  is  $-1.75\sigma$  or  $z = -1.75$ , we can now solve for  $x$  (the minimum guaranteed mileage):

$$z = \frac{x - 67,900}{2,050}$$

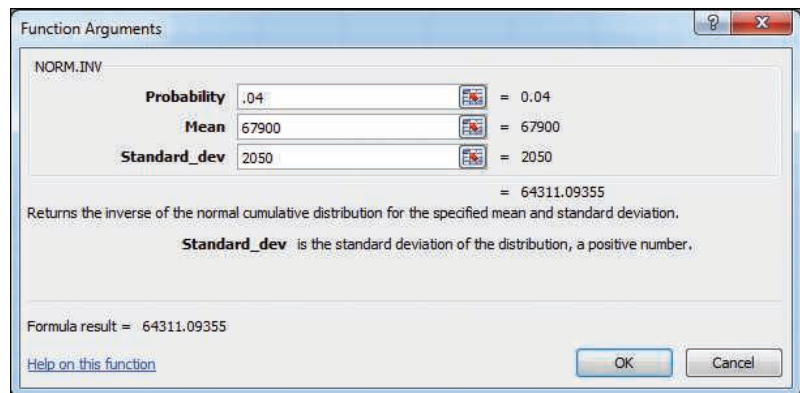
$$-1.75 = \frac{x - 67,900}{2,050}$$

$$-1.75(2,050) = x - 67,900$$

$$x = 67,900 - 1.75(2,050) = 64,312$$

So Layton can advertise that it will replace for free any tire that wears out before it reaches 64,312 miles, and the company will know that only 4% of the tires will be replaced under this plan.

Excel will also find the mileage value. See the following output. The necessary commands are given in the **Software Commands** in Appendix C.



Source: Microsoft Excel

## SELF-REVIEW 7-6



An analysis of the final test scores for Introduction to Business reveals the scores follow the normal probability distribution. The mean of the distribution is 75 and the standard deviation is 8. The professor wants to award an A to students whose score is in the highest 10%. What is the dividing point for those students who earn an A and those earning a B?

## EXERCISES

23. A normal distribution has a mean of 50 and a standard deviation of 4. Determine the value below which 95% of the observations will occur.
24. A normal distribution has a mean of 80 and a standard deviation of 14. Determine the value above which 80% of the values will occur.
25. Assume that the hourly cost to operate a commercial airplane follows the normal distribution with a mean of \$2,100 per hour and a standard deviation of \$250. What is the operating cost for the lowest 3% of the airplanes?
26. The SAT Reasoning Test is perhaps the most widely used standardized test for college admissions in the United States. Scores are based on a normal distribution with a mean of 1500 and a standard deviation of 300. Clinton College would like to offer an honors scholarship to students who score in the top 10% of this test. What is the minimum score that qualifies for the scholarship?
27. According to media research, the typical American listened to 195 hours of music in the last year. This is down from 290 hours 4 years earlier. Dick Trythall is a big country and western music fan. He listens to music while working around the house, reading, and riding in his truck. Assume the number of hours spent listening to music follows a normal probability distribution with a standard deviation of 8.5 hours.
  - a. If Dick is in the top 1% in terms of listening time, how many hours did he listen last year?
  - b. Assume that the distribution of times 4 years earlier also follows the normal probability distribution with a standard deviation of 8.5 hours. How many hours did the 1% who listen to the *least* music actually listen?
28. For the most recent year available, the mean annual cost to attend a private university in the United States was \$42,224. Assume the distribution of annual costs follows the normal probability distribution and the standard deviation is \$4,500. Ninety-five percent of all students at private universities pay less than what amount?
29. In economic theory, a “hurdle rate” is the minimum return that a person requires before he or she will make an investment. A research report says that annual returns from a specific class of common equities are distributed according to a normal distribution with a mean of 12% and a standard deviation of 18%. A stock screener would like to identify a hurdle rate such that only 1 in 20 equities is above that value. Where should the hurdle rate be set?
30. The manufacturer of a laser printer reports the mean number of pages a cartridge will print before it needs replacing is 12,200. The distribution of pages printed per cartridge closely follows the normal probability distribution and the standard deviation is 820 pages. The manufacturer wants to provide guidelines to potential customers as to how long they can expect a cartridge to last. How many pages should the manufacturer advertise for each cartridge if it wants to be correct 99% of the time?

## CHAPTER SUMMARY

- I. The uniform distribution is a continuous probability distribution with the following characteristics.
  - A. It is rectangular in shape.
  - B. The mean and the median are equal.
  - C. It is completely described by its minimum value  $a$  and its maximum value  $b$ .
  - D. It is described by the following equation for the region from  $a$  to  $b$ :

$$P(x) = \frac{1}{b - a} \quad (7-3)$$

E. The mean and standard deviation of a uniform distribution are computed as follows:

$$\mu = \frac{(a + b)}{2} \quad (7-1)$$

$$\sigma = \sqrt{\frac{(b - a)^2}{12}} \quad (7-2)$$

II. The normal probability distribution is a continuous distribution with the following characteristics.

- A. It is bell-shaped and has a single peak at the center of the distribution.
- B. The distribution is symmetric.
- C. It is asymptotic, meaning the curve approaches but never touches the  $X$ -axis.
- D. It is completely described by its mean and standard deviation.
- E. There is a family of normal probability distributions.
  1. Another normal probability distribution is created when either the mean or the standard deviation changes.
  2. The normal probability distribution is described by the following formula:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]} \quad (7-4)$$

III. The standard normal probability distribution is a particular normal distribution.

- A. It has a mean of 0 and a standard deviation of 1.
- B. Any normal probability distribution can be converted to the standard normal probability distribution by the following formula.

$$z = \frac{x - \mu}{\sigma} \quad (7-5)$$

C. By standardizing a normal probability distribution, we can report the distance of a value from the mean in units of the standard deviation.

## CHAPTER EXERCISES

31. The amount of cola in a 12-ounce can is uniformly distributed between 11.96 ounces and 12.05 ounces.
  - a. What is the mean amount per can?
  - b. What is the standard deviation amount per can?
  - c. What is the probability of selecting a can of cola and finding it has less than 12 ounces?
  - d. What is the probability of selecting a can of cola and finding it has more than 11.98 ounces?
  - e. What is the probability of selecting a can of cola and finding it has more than 11.00 ounces?
32. A tube of Listerine Tartar Control toothpaste contains 4.2 ounces. As people use the toothpaste, the amount remaining in any tube is random. Assume the amount of toothpaste remaining in the tube follows a uniform distribution. From this information, we can determine the following information about the amount remaining in a toothpaste tube without invading anyone's privacy.
  - a. How much toothpaste would you expect to be remaining in the tube?
  - b. What is the standard deviation of the amount remaining in the tube?
  - c. What is the likelihood there is less than 3.0 ounces remaining in the tube?
  - d. What is the probability there is more than 1.5 ounces remaining in the tube?
33. Many retail stores offer their own credit cards. At the time of the credit application, the customer is given a 10% discount on the purchase. The time required for the credit application process follows a uniform distribution with the times ranging from 4 minutes to 10 minutes.
  - a. What is the mean time for the application process?
  - b. What is the standard deviation of the process time?
  - c. What is the likelihood a particular application will take less than 6 minutes?
  - d. What is the likelihood an application will take more than 5 minutes?

- 34.** The time patrons at the Grande Dunes Hotel in the Bahamas spend waiting for an elevator follows a uniform distribution between 0 and 3.5 minutes.
- Show that the area under the curve is 1.00.
  - How long does the typical patron wait for elevator service?
  - What is the standard deviation of the waiting time?
  - What percent of the patrons wait for less than a minute?
  - What percent of the patrons wait more than 2 minutes?
- 35.** The net sales and the number of employees for aluminum fabricators with similar characteristics are organized into frequency distributions. Both are normally distributed. For the net sales, the mean is \$180 million and the standard deviation is \$25 million. For the number of employees, the mean is 1,500 and the standard deviation is 120. Clarion Fabricators had sales of \$170 million and 1,850 employees.
- Convert Clarion's sales and number of employees to  $z$  values.
  - Locate the two  $z$  values.
  - Compare Clarion's sales and number of employees with those of the other fabricators.
- 36.** The accounting department at Weston Materials Inc., a national manufacturer of unattached garages, reports that it takes two construction workers a mean of 32 hours and a standard deviation of 2 hours to erect the Red Barn model. Assume the assembly times follow the normal distribution.
- Determine the  $z$  values for 29 and 34 hours. What percent of the garages take between 32 hours and 34 hours to erect?
  - What percent of the garages take between 29 hours and 34 hours to erect?
  - What percent of the garages take 28.7 hours or less to erect?
  - Of the garages, 5% take how many hours or more to erect?
- 37.** Recently the United States Department of Agriculture issued a report (<http://www.cnpp.usda.gov/sites/default/files/CostofFoodMar2015.pdf>) indicating a family of four spent an average of about \$890 per month on food. Assume the distribution of food expenditures for a family of four follows the normal distribution, with a standard deviation of \$90 per month.
- What percent of the families spend more than \$430 but less than \$890 per month on food?
  - What percent of the families spend less than \$830 per month on food?
  - What percent spend between \$830 and \$1,000 per month on food?
  - What percent spend between \$900 and \$1,000 per month on food?
- 38.** A study of long-distance phone calls made from General Electric Corporate Headquarters in Fairfield, Connecticut, revealed the length of the calls, in minutes, follows the normal probability distribution. The mean length of time per call was 4.2 minutes and the standard deviation was 0.60 minute.
- What is the probability that calls last between 4.2 and 5 minutes?
  - What is the probability that calls last more than 5 minutes?
  - What is the probability that calls last between 5 and 6 minutes?
  - What is the probability that calls last between 4 and 6 minutes?
  - As part of her report to the president, the director of communications would like to report the length of the longest (in duration) 4% of the calls. What is this time?
- 39.** Shaver Manufacturing Inc. offers dental insurance to its employees. A recent study by the human resource director shows the annual cost per employee per year followed the normal probability distribution, with a mean of \$1,280 and a standard deviation of \$420 per year.
- What is the probability that annual dental expenses are more than \$1,500?
  - What is the probability that annual dental expenses are between \$1,500 and \$2,000?
  - Estimate the probability that an employee had no annual dental expenses.
  - What was the cost for the 10% of employees who incurred the highest dental expense?
- 40.** The annual commissions earned by sales representatives of Machine Products Inc., a manufacturer of light machinery, follow the normal probability distribution. The mean yearly amount earned is \$40,000 and the standard deviation is \$5,000.
- What percent of the sales representatives earn more than \$42,000 per year?
  - What percent of the sales representatives earn between \$32,000 and \$42,000?
  - What percent of the sales representatives earn between \$32,000 and \$35,000?

- d. The sales manager wants to award the sales representatives who earn the largest commissions a bonus of \$1,000. He can award a bonus to 20% of the representatives. What is the cutoff point between those who earn a bonus and those who do not?
41. According to the South Dakota Department of Health, the number of hours of TV viewing per week is higher among adult women than adult men. A recent study showed women spent an average of 34 hours per week watching TV and men, 29 hours per week. Assume that the distribution of hours watched follows the normal distribution for both groups and that the standard deviation among the women is 4.5 hours and is 5.1 hours for the men.
- What percent of the women watch TV less than 40 hours per week?
  - What percent of the men watch TV more than 25 hours per week?
  - How many hours of TV do the 1% of women who watch the most TV per week watch? Find the comparable value for the men.
42. According to a government study, among adults in the 25- to 34-year age group, the mean amount spent per year on reading and entertainment is \$1,994. Assume that the distribution of the amounts spent follows the normal distribution with a standard deviation of \$450.
- What percent of the adults spend more than \$2,500 per year on reading and entertainment?
  - What percent spend between \$2,500 and \$3,000 per year on reading and entertainment?
  - What percent spend less than \$1,000 per year on reading and entertainment?
43. Management at Gordon Electronics is considering adopting a bonus system to increase production. One suggestion is to pay a bonus on the highest 5% of production, based on past experience. Past records indicate weekly production follows the normal distribution. The mean of this distribution is 4,000 units per week and the standard deviation is 60 units per week. If the bonus is paid on the upper 5% of production, the bonus will be paid on how many units or more?
44. Fast Service Truck Lines uses the Ford Super Duty F-750 exclusively. Management made a study of the maintenance costs and determined the number of miles traveled during the year followed the normal distribution. The mean of the distribution was 60,000 miles and the standard deviation was 2,000 miles.
- What percent of the Ford Super Duty F-750s logged 65,200 miles or more?
  - What percent of the trucks logged more than 57,060 but less than 58,280 miles?
  - What percent of the Fords traveled 62,000 miles or less during the year?
  - Is it reasonable to conclude that any of the trucks were driven more than 70,000 miles? Explain.
45. Best Electronics Inc. offers a “no hassle” returns policy. The daily number of customers returning items follows the normal distribution. The mean number of customers returning items is 10.3 per day and the standard deviation is 2.25 per day.
- For any day, what is the probability that eight or fewer customers returned items?
  - For any day, what is the probability that the number of customers returning items is between 12 and 14?
  - Is there any chance of a day with no customer returns?
46. The funds dispensed at the ATM machine located near the checkout line at the Kroger’s in Union, Kentucky, follows a normal probability distribution with a mean of \$4,200 per day and a standard deviation of \$720 per day. The machine is programmed to notify the nearby bank if the amount dispensed is very low (less than \$2,500) or very high (more than \$6,000).
- What percent of the days will the bank be notified because the amount dispensed is very low?
  - What percent of the time will the bank be notified because the amount dispensed is high?
  - What percent of the time will the bank not be notified regarding the amount of funds dispensed?
47. The weights of canned hams processed at Henline Ham Company follow the normal distribution, with a mean of 9.20 pounds and a standard deviation of 0.25 pound. The label weight is given as 9.00 pounds.
- What proportion of the hams actually weigh less than the amount claimed on the label?



- b. The owner, Glen Henline, is considering two proposals to reduce the proportion of hams below label weight. He can increase the mean weight to 9.25 and leave the standard deviation the same, or he can leave the mean weight at 9.20 and reduce the standard deviation from 0.25 pound to 0.15. Which change would you recommend?
48. A recent Gallup study ([http://www.gallup.com/poll/175286/hour-workweek-actually-longer-seven-hours.aspx?g\\_source=polls+work+hours&g\\_medium=search&g\\_campaign=tiles](http://www.gallup.com/poll/175286/hour-workweek-actually-longer-seven-hours.aspx?g_source=polls+work+hours&g_medium=search&g_campaign=tiles)) found the typical American works an average of 46.7 hour per week. The study did not report the shape of the distribution of hours worked or the standard deviation. It did, however, indicate that 40% of the workers worked less than 40 hours a week and that 18 percent worked more than 60 hours.
- If we assume that the distribution of hours worked is normally distributed, and knowing 40% of the workers worked less than 40 hours, find the standard deviation of the distribution.
  - If we assume that the distribution of hours worked is normally distributed and 18% of the workers worked more than 60 hours, find the standard deviation of the distribution.
  - Compare the standard deviations computed in parts *a* and *b*. Is the assumption that the distribution of hours worked is approximately normal reasonable? Why?
49. Most four-year automobile leases allow up to 60,000 miles. If the lessee goes beyond this amount, a penalty of 20 cents per mile is added to the lease cost. Suppose the distribution of miles driven on four-year leases follows the normal distribution. The mean is 52,000 miles and the standard deviation is 5,000 miles.
- What percent of the leases will yield a penalty because of excess mileage?
  - If the automobile company wanted to change the terms of the lease so that 25% of the leases went over the limit, where should the new upper limit be set?
  - One definition of a low-mileage car is one that is 4 years old and has been driven less than 45,000 miles. What percent of the cars returned are considered low-mileage?
50. The price of shares of Bank of Florida at the end of trading each day for the last year followed the normal distribution. Assume there were 240 trading days in the year. The mean price was \$42.00 per share and the standard deviation was \$2.25 per share.
- What is the probability that the end-of-day trading price is over \$45.00? Estimate the number of days in a year when the trading price finished above \$45.00.
  - What percent of the days was the price between \$38.00 and \$40.00?
  - What is the minimum share price for the top 15% of end-of-day trading prices?
51. The annual sales of romance novels follow the normal distribution. However, the mean and the standard deviation are unknown. Forty percent of the time, sales are more than 470,000, and 10% of the time, sales are more than 500,000. What are the mean and the standard deviation?
52. In establishing warranties on HDTVs, the manufacturer wants to set the limits so that few will need repair at the manufacturer's expense. On the other hand, the warranty period must be long enough to make the purchase attractive to the buyer. For a new HDTV, the mean number of months until repairs are needed is 36.84 with a standard deviation of 3.34 months. Where should the warranty limits be set so that only 10% of the HDTVs need repairs at the manufacturer's expense?

## DATA ANALYTICS

(The data for these exercises are available in Connect.)

53. Refer to the North Valley Real Estate data, which report information on homes sold during the last year.
- The mean selling price (in \$ thousands) of the homes was computed earlier to be \$357.0, with a standard deviation of \$160.7. Use the normal distribution to estimate the percentage of homes selling for more than \$500,000. Compare this to the actual results. Is price normally distributed? Try another test. If price is normally distributed, how many homes should have a price greater than the mean? Compare this to the actual number of homes. Construct a frequency distribution of price. What do you observe?

- b. The mean days on the market is 30 with a standard deviation of 10 days. Use the normal distribution to estimate the number of homes on the market more than 24 days. Compare this to the actual results. Try another test. If days on the market is normally distributed, how many homes should be on the market more than the mean number of days? Compare this to the actual number of homes. Does the normal distribution yield a good approximation of the actual results? Create a frequency distribution of days on the market. What do you observe?
54. Refer to the Baseball 2016 data, which report information on the 30 Major League Baseball teams for the 2016 season.
- a. The mean attendance per team for the season was 2.439 million, with a standard deviation of 0.618 million. Use the normal distribution to estimate the number of teams with attendance of more than 3.5 million. Compare that estimate with the actual number. Comment on the accuracy of your estimate.
- b. The mean team salary was \$121 million, with a standard deviation of \$40.0 million. Use the normal distribution to estimate the number of teams with a team salary of more than \$100 million. Compare that estimate with the actual number. Comment on the accuracy of the estimate.
55. Refer to the Lincolnville School District bus data.
- a. Refer to the maintenance cost variable. The mean maintenance cost for last year is \$4,552 with a standard deviation of \$2332. Estimate the number of buses with a maintenance cost of more than \$6,000. Compare that with the actual number. Create a frequency distribution of maintenance cost. Is the distribution normally distributed?
- b. Refer to the variable on the number of miles driven since the last maintenance. The mean is 11,121 and the standard deviation is 617 miles. Estimate the number of buses traveling more than 11,500 miles since the last maintenance. Compare that number with the actual value. Create a frequency distribution of miles since maintenance cost. Is the distribution normally distributed?

## PRACTICE TEST

### Part 1—Objective

- For a continuous probability distribution, the total area under the curve is equal to \_\_\_\_\_.
- For a uniform distribution that ranges from 10 to 20, how many values can be in that range? (1, 10, 100, infinite—pick one) \_\_\_\_\_.
- Which of the following is NOT a characteristic of the normal distribution? (bell-shaped, symmetrical, discrete, asymptotic—pick one) \_\_\_\_\_.
- For a normal distribution, what is true about the mean and median? (always equal, the mean is twice the median, the mean and median are equal to the standard deviation, none of these is true—pick one) \_\_\_\_\_.
- How many normal distributions are there? (1, 10, 30, infinite—pick one) \_\_\_\_\_.
- How many standard normal distributions are there? (1, 10, 30, infinite—pick one) \_\_\_\_\_.
- The signed difference between a selected value and the mean divided by the standard deviation is called a \_\_\_\_\_. (z score, z value, standardized value, all of these—pick one)
- What is the probability of a z value between 0 and  $-0.76$ ? \_\_\_\_\_
- What is the probability of a z value between  $-2.03$  and  $1.76$ ? \_\_\_\_\_
- What is the probability of a z value between  $-1.86$  and  $-1.43$ ? \_\_\_\_\_

### Part 2—Problem

- The IRS reports that the mean refund for a particular group of taxpayers was \$1,600. The distribution of tax refunds follows a normal distribution with a standard deviation of \$850.
  - What percentage of the refunds are between \$1,600 and \$2,000?
  - What percentage of the refunds are between \$900 and \$2,000?
  - What percentage of the refunds are between \$1,800 and \$2,000?
  - Ninety-five percent of the refunds are for less than what amount?

# 8

# Sampling Methods and the Central Limit Theorem



©August\_0802/Shutterstock

- ▲ **THE NIKE** annual report says that the average American buys 6.5 pairs of sports shoes per year. Suppose a sample of 81 customers is surveyed and the population standard deviation of sports shoes purchased per year is 2.1. What is the standard error of the mean in this experiment? (See Exercise 45 and **L08-4**.)

## LEARNING OBJECTIVES

---

*When you have completed this chapter, you will be able to:*

- L08-1** Explain why populations are sampled and describe four methods to sample a population.
- L08-2** Define sampling error.
- L08-3** Explain the sampling distribution of the sample mean.
- L08-4** Recite the central limit theorem and define the mean and standard error of the sampling distribution of the sample mean.
- L08-5** Apply the central limit theorem to calculate probabilities.

## INTRODUCTION

Chapters 2 through 4 emphasize techniques to describe data. To illustrate these techniques, we organize the profits for the sale of 180 vehicles by the four dealers included in the Applewood Auto Group into a frequency distribution and compute measures of location and dispersion. Such measures as the mean and the standard deviation describe the typical profit and the spread in the profits. In these chapters, the emphasis is on describing the distribution of the data. That is, we describe something that has already happened.

In Chapter 5, we begin to lay the foundation for statistical inference with the study of probability. Recall that in statistical inference our goal is to determine something about a *population* based only on the *sample*. The population is the entire group of individuals or objects under consideration, and the sample is a part or subset of that population. Chapter 6 extends the probability concepts by describing two discrete probability distributions: the binomial and the Poisson. Chapter 7 describes two continuous probability distributions: the uniform and normal. Probability distributions encompass all possible outcomes of an experiment and the probability associated with each outcome. We use probability distributions to evaluate the likelihood something will occur in the future.

This chapter begins our study of sampling. Sampling is a process of selecting items from a population so we can use this information to make judgments or inferences about the population. We begin this chapter by discussing methods of selecting a sample from a population. Next, we construct a distribution of the sample mean to understand how the sample means tend to cluster around the population mean. Finally, we show that for any population the shape of this sampling distribution tends to follow the normal probability distribution.

### LO8-1

Explain why populations are sampled and describe four methods to sample a population.

## SAMPLING METHODS

In Chapter 1, we said the purpose of inferential statistics is to find something about a population based on a sample. A sample is a portion or part of the population of interest. In many cases, sampling is more feasible than studying the entire population. In this section, we discuss the reasons for sampling and then several methods for selecting a sample.

### Reasons to Sample

When studying characteristics of a population, there are many practical reasons why we prefer to select portions or samples of a population to observe and measure. Here are some of the reasons for sampling:

1. **To contact the whole population would be time-consuming.** A candidate for a national office may wish to determine her chances for election. A sample poll using the regular staff and field interviews of a professional polling firm would take only 1 or 2 days. Using the same staff and interviewers and working 7 days a week, it would take nearly 200 years to contact all the voting population! Even if a large staff of interviewers could be assembled, the benefit of contacting all of the voters would probably not be worth the time.
2. **The cost of studying all the items in a population may be prohibitive.** Public opinion polls and consumer testing organizations, such as Harris Interactive Inc., CBS News Polls, and Zogby Analytics, usually contact fewer than 2,000 of the nearly 60 million families in the United States. One consumer panel-type organization charges \$40,000 to mail samples and tabulate responses to test a product (such as breakfast cereal, cat food, or perfume). The same product test using all 60 million families would be too expensive to be worthwhile.
3. **The physical impossibility of checking all items in the population.** Some populations are infinite. It would be impossible to check all the water in Lake Erie for bacterial levels, so we select samples at various locations. The populations of fish, birds, snakes, deer, and the like are large and are constantly moving, being



©David Epperson/Getty Images RF

born, and dying. Instead of even attempting to count all the ducks in Canada or all the fish in Lake Pontchartrain, we make estimates using various techniques—such as counting all the ducks on a pond selected at random, tracking fish catches, or netting fish at predetermined places in a lake.

4. **The destructive nature of some tests.** If the wine tasters at the Sutter Home Winery in California drank all the wine to evaluate the vintage, they would consume the entire crop, and none would be available for sale. In the area of industrial production, steel plates, wires, and similar products must have a certain minimum tensile strength. To ensure that the product meets the minimum standard, the quality assurance department selects a sample from the current production. Each piece is stretched until it breaks and the breaking point (usually measured in pounds per square inch)

recorded. Obviously, if all the wire or all the plates were tested for tensile strength, none would be available for sale or use. For the same reason, only a few seeds are tested for germination by Burpee Seeds Inc. prior to the planting season.

5. **The sample results are adequate.** Even if funds were available, it is doubtful the additional accuracy of a 100% sample—that is, studying the entire population—is essential in most problems. For example, the federal government uses a sample of grocery stores scattered throughout the United States to determine the monthly index of food prices. The prices of bread, beans, milk, and other major food items are included in the index. It is unlikely that the inclusion of all grocery stores in the United States would significantly affect the index because the prices of milk, bread, and other major foods usually do not vary by more than a few cents from one chain store to another.

## Simple Random Sampling

The most widely used sampling method is a **simple random sampling**.

**SIMPLE RANDOM SAMPLE** A sample selected so that each item or person in the population has the same chance of being included.

To illustrate the selection process for a simple random sample, suppose the population of interest is the 750 Major League Baseball players on the active rosters of the 30 teams at the end of the 2017 season. The president of the players' union wishes to form a committee of 10 players to study the issue of concussions. One way of ensuring that every player in the population has the same chance of being chosen to serve on the Concussion Committee is to write each name of the 750 players on a slip of paper and place all the slips of paper in a box. After the slips of paper have been thoroughly mixed, the first selection is made by drawing a slip of paper from the box identifying the first player. The slip of paper is not returned to the box. This process is repeated nine more times to form the committee. (Note that the probability of each selection does increase slightly because the slip is not replaced. However, the differences are very small because the population is 750. The probability of each selection is about 0.0013, rounded to four decimal places.)

Of course, the process of writing all the players' names on a slip of paper is very time-consuming. A more convenient method of selecting a random sample is to use a **table of random numbers** such as the one in Appendix B.4. In this case the union president would prepare a list of all 750 players and number each of the players from 1 to 750 with a computer application. Using a table of random numbers, we would randomly pick a starting place in the table, and then select 10 three-digit numbers between 001 and 750. A computer can also generate random numbers. These numbers would correspond with the 10 players in the list that will be asked to participate on the committee. As the name *simple random sampling* implies, the probability of selecting any number between 001 and 750 is the same. Thus, the probability of selecting the player assigned the number 131 is

**STATISTICS IN ACTION**

To ensure that an unbiased, representative sample is selected from a population, lists of random numbers are needed. In 1927, L. Tippett published the first book of random numbers. In 1938, R. A. Fisher and F. Yates published 15,000 random digits generated using two decks of cards. In 1955, RAND Corporation published a million random digits, generated by the random frequency pulses of an electronic roulette wheel. Since then, computer programs have been developed for generating digits that are “almost” random and hence are called *pseudo-random*. The question of whether a computer program can be used to generate numbers that are truly random remains a debatable issue.

the same as the probability of selecting player 722 or player 382. Using random numbers to select players for the committee removes any bias from the selection process.

The following example shows how to select random numbers using a portion of a random number table illustrated below. First, we choose a starting point in the table. One way of selecting the starting point is to close your eyes and point at a number in the table. Any starting point will do. Another way is to randomly pick a column and row. Suppose the time is 3:04. Using the hour, three o’clock, pick the third column and then, using the minutes, four, move down to the fourth row of numbers. The number is 03759. Because there are only 750 players, we will use the first three digits of a five-digit random number. Thus, 037 is the number of the first player to be a member of the sample. To continue selecting players, we could move in any direction. Suppose we move right. The first three digits of the number to the right of 03759 are 447. Player number 447 is the second player selected to be on the committee. The next three-digit number to the right is 961. You skip 961 as well as the next number 784 because there are only 750 players. The third player selected is number 189. We continue this process until we have 10 players.

|       |       |                |               |       |       |              |
|-------|-------|----------------|---------------|-------|-------|--------------|
| 50525 | 57454 | 28455          | 68226         | 34656 | 38884 | 39018        |
| 72507 | 53380 | 53827          | 42486         | 54465 | 71819 | 91199        |
| 34986 | 74297 | 00144          | 38676         | 89967 | 98869 | 39744        |
| 68851 | 27305 | 03759          | 44723         | 96108 | 78489 | 18910        |
| 06738 | 62879 | 03910          | 17350         | 49169 | 03850 | 18910        |
| 11448 | 10734 | 05837          | 24397         | 10420 | 16712 | 94496        |
|       |       | Starting point | Second player |       |       | Third player |

Statistical packages such as Minitab and spreadsheet packages such as Excel have software that will select a simple random sample. The following example/solution uses Excel to select a random sample from a list of the data.

**EXAMPLE**

Jane and Joe Miley operate the Foxtrot Inn, a bed and breakfast in Tryon, North Carolina. There are eight rooms available for rent at this B&B. For each day of June 2017, the number of rooms rented is listed. Use Excel to select a sample of five nights during the month of June.

| June | Rentals | June | Rentals | June | Rentals |
|------|---------|------|---------|------|---------|
| 1    | 0       | 11   | 3       | 21   | 3       |
| 2    | 2       | 12   | 4       | 22   | 2       |
| 3    | 3       | 13   | 4       | 23   | 3       |
| 4    | 2       | 14   | 4       | 24   | 6       |
| 5    | 3       | 15   | 7       | 25   | 0       |
| 6    | 4       | 16   | 0       | 26   | 4       |
| 7    | 2       | 17   | 5       | 27   | 1       |
| 8    | 3       | 18   | 3       | 28   | 1       |
| 9    | 4       | 19   | 6       | 29   | 3       |
| 10   | 7       | 20   | 2       | 30   | 3       |

**SOLUTION**

Excel will select a random sample and report the results. Sampling is done *with* replacement, so it is possible that the same day may appear more than once in the sample. On the first sampled date, four of the eight rooms were rented. On the second sampled date

in June, seven rooms were rented. The information is reported in column D of the Excel spreadsheet. The steps are listed in the **Software Commands** in Appendix C.

|    | A           | B       | C | D      |
|----|-------------|---------|---|--------|
| 1  | Day of June | Rentals |   | Sample |
| 2  | 1           | 0       |   | 4      |
| 3  | 2           | 2       |   | 7      |
| 4  | 3           | 3       |   | 4      |
| 5  | 4           | 2       |   | 3      |
| 6  | 5           | 3       |   | 1      |
| 7  | 6           | 4       |   |        |
| 8  | 7           | 2       |   |        |
| 9  | 8           | 3       |   |        |
| 10 | 9           | 4       |   |        |
| 11 | 10          | 7       |   |        |
| 12 | 11          | 3       |   |        |
| 13 | 12          | 4       |   |        |
| 14 | 13          | 4       |   |        |
| 15 | 14          | 4       |   |        |

Source: Microsoft Excel

## SELF-REVIEW 8-1



The following roster lists the students enrolled in an introductory course in business statistics. Three students will be randomly selected and asked questions about course content and method of instruction.

- The numbers 00 through 45 are handwritten on slips of paper and placed in a bowl. The three numbers selected are 31, 7, and 25. Which students are in the sample?
- Now use the table of random numbers, Appendix B.4, to select your own sample.
- What would you do if you encountered the number 59 in the table of random digits?

| STAT 264 BUSINESS STATISTICS                           |                          |            |               |                           |            |
|--|--------------------------|------------|---------------|---------------------------|------------|
| 9:00 AM - 9:50 AM MW; 118 CARLSON HALL; PROFESSOR LIND |                          |            |               |                           |            |
| RANDOM NUMBER  | NAME                     | CLASS RANK | RANDOM NUMBER | NAME                      | CLASS RANK |
| 00   | ANDERSON, RAYMOND        | SO         | 23            | MEDLEY, CHERYL ANN        | SO         |
| 01   | ANGER, CHERYL RENEE      | SO         | 24            | MITCHELL, GREG R          | FR         |
| 02   | BALL, CLAIRE JEANETTE    | FR         | 25            | MOLTER, KRISTI MARIE      | SO         |
| 03   | BERRY, CHRISTOPHER G     | FR         | 26            | MULCAHY, STEPHEN ROBERT   | SO         |
| 04   | BOBAK, JAMES PATRICK     | SO         | 27            | NICHOLAS, ROBERT CHARLES  | JR         |
| 05   | BRIGHT, M. STARR         | JR         | 28            | NICKENS, VIRGINIA         | SO         |
| 06   | CHONTOS, PAUL JOSEPH     | SO         | 29            | PENNYWITT, SEAN PATRICK   | SO         |
| 07   | DETLEY, BRIAN HANS       | JR         | 30            | POTEAU, KRIS E            | JR         |
| 08   | DUDAS, VIOLA             | SO         | 31            | PRICE, MARY LYNETTE       | SO         |
| 09   | DULBS, RICHARD ZALFA     | JR         | 32            | RISTAS, JAMES             | SR         |
| 10   | EDINGER, SUSAN KEE       | SR         | 33            | SAGER, ANNE MARIE         | SO         |
| 11   | FINK, FRANK JAMES        | SR         | 34            | SMILLIE, HEATHER MICHELLE | SO         |
| 12   | FRANCIS, JAMES P         | JR         | 35            | SNYDER, LEISHA KAY        | SR         |
| 13   | GAGHEN, PAMELA LYNN      | JR         | 36            | STAHL, MARIA TASHERY      | SO         |
| 14   | GOULD, ROBYN KAY         | SO         | 37            | ST. JOHN, AMY J           | SO         |
| 15   | GROSENBACHER, SCOTT ALAN | SO         | 38            | STURDEVANT, RICHARD K     | SO         |
| 16   | HEETFIELD, DIANE MARIE   | SO         | 39            | SWETYE, LYNN MICHELE      | SO         |
| 17   | KABAT, JAMES DAVID       | JR         | 40            | WALASINSKI, MICHAEL       | SO         |
| 18   | KEMP, LISA ADRIANE       | FR         | 41            | WALKER, DIANE ELAINE      | SO         |
| 19   | KILLION, MICHELLE A      | SO         | 42            | WARNOCK, JENNIFER MARY    | SO         |
| 20   | KOPERSKI, MARY ELLEN     | SO         | 43            | WILLIAMS, WENDY A         | SO         |
| 21   | KOPP, BRIDGETTE ANN      | SO         | 44            | YAP, HOCK BAN             | SO         |
| 22   | LEHMANN, KRISTINA MARIE  | JR         | 45            | YODER, ARLAN JAY          | JR         |

**STATISTICS IN ACTION**

Random and unbiased sampling methods are extremely important to make valid statistical inferences. In 1936, the *Literary Digest* conducted a straw vote to predict the outcome of the presidential race between Franklin Roosevelt and Alfred Landon. Ten million ballots in the form of returnable postcards were sent to addresses taken from *Literary Digest* subscribers, telephone directories, and automobile registrations. In 1936, not many people could afford a telephone or an automobile. Thus, the population that was sampled did not represent the population of voters. A second problem was with the non-responses. More than 10 million people were sent surveys, and more than 2.3 million responded. However, no attempt was made to see whether those responding represented a cross-section of all the voters. The sample information predicted Landon would win with 57% of the vote and Roosevelt would have 43%. On Election Day, Roosevelt won with 61% of the vote. Landon had 39%. In the mid-1930s, people who had telephones and drove automobiles clearly did not represent American voters!

## Systematic Random Sampling

The simple random sampling procedure is awkward in some research situations. For example, Stood's Grocery Market needs to sample their customers to study the length of time customers spend in the store. Simple random sampling is not an effective method. Practically, we do not have a list of customers, so assigning random numbers to customers is impossible. Instead, we can use **systematic random sampling** to select a representative sample. Using this method for Stood's Grocery Market, we decide to select 100 customers over 4 days, Monday through Thursday. We will select 25 customers a day and begin the sampling at different times each day: 8 a.m., 11 a.m., 4 p.m., and 7 p.m. We write the 4 times and 4 days on slips of paper and put them in two hats—one hat for the days and the other hat for the times. We select one slip from each hat. This ensures that the time of day is randomly assigned for each day. Suppose we selected 4 p.m. for the starting time on Monday. Next we select a random number between 1 and 10; it is 6. Our selection process begins on Monday at 4 p.m. by selecting the sixth customer to enter the store. Then, we select every 10th (16th, 26th, 36th) customer until we reach the goal of 25 customers. For each of these sampled customers, we measure the length of time the customer spends in the store.

**SYSTEMATIC RANDOM SAMPLE** A random starting point is selected, and then every  $k$ th member of the population is selected.

Simple random sampling is used in the selection of the days, the times, and the starting point. But the systematic procedure is used to select the actual customer.

Before using systematic random sampling, we should carefully observe the physical order of the population. When the physical order is related to the population characteristic, then systematic random sampling should not be used because the sample could be biased. For example, if we wanted to audit the invoices in a file drawer that were ordered in increasing dollar amounts, systematic random sampling would not guarantee an unbiased random sample. Other sampling methods should be used.

## Stratified Random Sampling

When a population can be clearly divided into groups based on some characteristic, we may use **stratified random sampling**. It guarantees each group is represented in the sample. The groups are called **strata**. For example, college students can be grouped as full time or part time; as male or female; or as freshman, sophomore, junior, or senior. Usually the strata are formed based on members' shared attributes or characteristics. A random sample from each stratum is taken in a number proportional to the stratum's size when compared to the population. Once the strata are defined, we apply simple random sampling within each group or stratum to collect the sample.

**STRATIFIED RANDOM SAMPLE** A population is divided into subgroups, called strata, and a sample is randomly selected from each stratum.

For instance, we might study the advertising expenditures for the 352 largest companies in the United States. The objective of the study is to determine whether firms with high returns on equity (a measure of profitability) spend more on advertising than firms with low returns on equity. To make sure the sample is a fair representation of the 352 companies, the companies are grouped on percent return on equity. Table 8–1 shows the strata and the relative frequencies. If simple random sampling is used, observe that firms in the 3rd and 4th strata have a high chance of selection (probability of 0.87) while firms in the other strata have a small chance of selection (probability of 0.13). We might not select any firms in stratum 1 or 5 *simply by chance*. However, stratified random sampling will guarantee that at least one firm in each of strata 1 and



**TABLE 8–1** Number Selected for a Proportional Stratified Random Sample

| Stratum | Profitability (return on equity) | Number of Firms | Relative Frequency | Number Sampled |
|---------|----------------------------------|-----------------|--------------------|----------------|
| 1       | 30% and over                     | 8               | 0.02               | 1*             |
| 2       | 20 up to 30%                     | 35              | 0.10               | 5*             |
| 3       | 10 up to 20%                     | 189             | 0.54               | 27             |
| 4       | 0 up to 10%                      | 115             | 0.33               | 16             |
| 5       | Deficit                          | 5               | 0.01               | 1              |
| Total   |                                  | 352             | 1.00               | 50             |

\*0.02 of 50 = 1, 0.10 of 50 = 5, etc.

5 is represented in the sample. Let’s say that 50 firms are selected for intensive study. Then based on probability, one firm, or  $(0.02)(50)$ , should be randomly selected from stratum 1. We would randomly select five, or  $(0.10)(50)$ , firms from stratum 2. In this case, the number of firms sampled from each stratum is proportional to the stratum’s relative frequency in the population. Stratified sampling has the advantage, in some cases, of more accurately reflecting the characteristics of the population than does simple random or systematic random sampling.

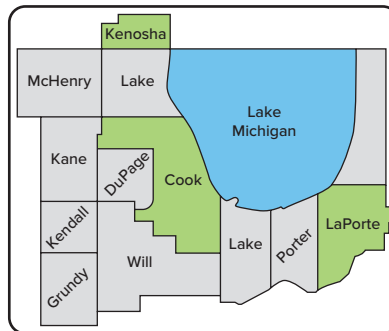
### Cluster Sampling

Another common type of sampling is **cluster sampling**. It is often employed to reduce the cost of sampling a population scattered over a large geographic area.

**CLUSTER SAMPLING** A population is divided into clusters using naturally occurring geographic or other boundaries. Then, clusters are randomly selected and a sample is collected by randomly selecting from each cluster.

Suppose you want to determine the views of residents in the greater Chicago, Illinois, metropolitan area about state and federal environmental protection policies. Selecting a random sample of residents in this region and personally contacting each one would be time-consuming and very expensive. Instead, you could employ cluster sampling by subdividing the region into small units, perhaps by counties. These are often called *primary units*.

There are 12 counties in the greater Chicago metropolitan area. Suppose you randomly select three counties. The three chosen are LaPorte, Cook, and Kenosha (see Chart 8–1 below). Next, you select a random sample of the residents in each of these counties and interview them. This is also referred to as sampling through an *intermediate unit*. In this case, the intermediate unit is the county. (Note that this is a combination of cluster sampling and simple random sampling.)



**CHART 8–1** The Counties of the Greater Chicago, Illinois, Metropolitan Area

The discussion of sampling methods in the preceding sections did not include all the sampling methods available to a researcher. Should you become involved in a major research project in marketing, finance, accounting, or other areas, you would need to consult books devoted solely to sample theory and sample design.

## SELF-REVIEW 8-2



Refer to Self-Review 8-1 and the class roster on page 214. Suppose a systematic random sample will select every ninth student enrolled in the class. Initially, the fourth student on the list was selected at random. That student is numbered 03. Remembering that the random numbers start with 00, which students will be chosen to be members of the sample?

## EXERCISES

- The following is a list of 24 Marco's Pizza stores in Lucas County. The stores are identified by numbering them 00 through 23. Also noted is whether the store is corporate-owned (C) or manager-owned (M). A sample of four locations is to be selected and inspected for customer convenience, safety, cleanliness, and other features.

| ID No. | Address           | Type | ID No. | Address              | Type |
|--------|-------------------|------|--------|----------------------|------|
| 00     | 2607 Starr Av     | C    | 12     | 2040 Ottawa River Rd | C    |
| 01     | 309 W Alexis Rd   | C    | 13     | 2116 N Reynolds Rd   | C    |
| 02     | 2652 W Central Av | C    | 14     | 3678 Rugby Dr        | C    |
| 03     | 630 Dixie Hwy     | M    | 15     | 1419 South Av        | C    |
| 04     | 3510 Dorr St      | C    | 16     | 1234 W Sylvania Av   | C    |
| 05     | 5055 Glendale Av  | C    | 17     | 4624 Woodville Rd    | M    |
| 06     | 3382 Lagrange St  | M    | 18     | 5155 S Main St       | M    |
| 07     | 2525 W Laskey Rd  | C    | 19     | 106 E Airport Hwy    | C    |
| 08     | 303 Louisiana Av  | C    | 20     | 6725 W Central Av    | M    |
| 09     | 149 Main St       | C    | 21     | 4252 Monroe St       | C    |
| 10     | 835 S McCord Rd   | M    | 22     | 2036 Woodville Rd    | C    |
| 11     | 3501 Monroe St    | M    | 23     | 1316 Michigan Av     | M    |

- The random numbers selected are 08, 18, 11, 54, 02, 41, and 54. Which stores are selected?
  - Use the table of random numbers to select your own sample of locations.
  - Using systematic random sampling, every seventh location is selected starting with the third store in the list. Which locations will be included in the sample?
  - Using stratified random sampling, select three locations. Two should be corporate-owned and one should be manager-owned.
- The following is a list of 29 hospitals in the Cincinnati, Ohio, and Northern Kentucky region. Each hospital is assigned a number, 00 through 28. The hospitals are classified by type, either a general medical/surgical hospital (M/S) or a specialty hospital (S). We are interested in estimating the average number of full- and part-time nurses employed in the area hospitals.

| ID Number | Name                     | Address  | Type | ID Number | Name                    | Address                                   | Type |
|-----------|--------------------------|--|------|-----------|-------------------------|---|------|
| 00        | Bethesda North           | 10500 Montgomery Rd. Cincinnati, Ohio 45242    | M/S  | 04        | Mercy Hospital-Hamilton | 100 Riverfront Plaza Hamilton, Ohio 45011 | M/S  |
| 01        | Ft. Hamilton-Hughes      | 630 Eaton Avenue Hamilton, Ohio 45013          | M/S  | 05        | Middletown Regional     | 105 McKnight Drive Middletown, Ohio 45044 | M/S  |
| 02        | Jewish Hospital-Kenwood  | 4700 East Galbraith Rd. Cincinnati, Ohio 45236 | M/S  | 06        | Clermont Mercy Hospital | 3000 Hospital Drive Batavia, Ohio 45103   | M/S  |
| 03        | Mercy Hospital-Fairfield | 3000 Mack Road Fairfield, Ohio 45014           | M/S  | 07        | Mercy Hospital-Anderson | 7500 State Road Cincinnati, Ohio 45255    | M/S  |

| ID Number | Name                                     | Address  | Type | ID Number | Name   | Address   | Type |
|-----------|--|--|------|-----------|--|---|------|
| 08        | Bethesda Oak Hospital                    | 619 Oak Street<br>Cincinnati, Ohio 45206         | M/S  | 19        | St. Luke's Hospital West                             | 7380 Turfway Drive<br>Florence, Kentucky 41075      | M/S  |
| 09        | Children's Hospital Medical Center       | 3333 Burnet Avenue<br>Cincinnati, Ohio 45229     | M/S  | 20        | St. Luke's Hospital East                             | 85 North Grand Avenue<br>Ft. Thomas, Kentucky 41042 | M/S  |
| 10        | Christ Hospital                          | 2139 Auburn Avenue<br>Cincinnati, Ohio 45219     | M/S  | 21        | Care Unit Hospital                                   | 3156 Glenmore Avenue<br>Cincinnati, Ohio 45211      | S    |
| 11        | Deaconess Hospital                       | 311 Straight Street<br>Cincinnati, Ohio 45219    | M/S  | 22        | Emerson Behavioral Science                           | 2446 Kipling Avenue<br>Cincinnati, Ohio 45239       | S    |
| 12        | Good Samaritan Hospital                  | 375 Dixmyth Avenue<br>Cincinnati, Ohio 45220     | M/S  | 23        | Pauline Warfield Lewis Center for Psychiatric Treat. | 1101 Summit Road<br>Cincinnati, Ohio 45237          | S    |
| 13        | Jewish Hospital                          | 3200 Burnet Avenue<br>Cincinnati, Ohio 45229     | M/S  | 24        | Children's Psychiatric No. Kentucky                  | 502 Farrell Drive<br>Covington, Kentucky 41011      | S    |
| 14        | University Hospital                      | 234 Goodman Street<br>Cincinnati, Ohio 45267     | M/S  | 25        | Drake Center Rehab—Long Term                         | 151 W. Galbraith Road<br>Cincinnati, Ohio 45216     | S    |
| 15        | Providence Hospital                      | 2446 Kipling Avenue<br>Cincinnati, Ohio 45239    | M/S  | 26        | No. Kentucky Rehab Hospital—Short Term               | 201 Medical Village<br>Edgewood, Kentucky           | S    |
| 16        | St. Francis—St. George Hospital          | 3131 Queen City Avenue<br>Cincinnati, Ohio 45238 | M/S  | 27        | Shriners Burns Institute                             | 3229 Burnet Avenue<br>Cincinnati, Ohio 45229        | S    |
| 17        | St. Elizabeth Medical Center, North Unit | 401 E. 20th Street<br>Covington, Kentucky 41014  | M/S  | 28        | VA Medical Center                                    | 3200 Vine Street<br>Cincinnati, Ohio 45220          | S    |
| 18        | St. Elizabeth Medical Center, South Unit | One Medical Village<br>Edgewood, Kentucky 41017  | M/S  |           |  |   |      |

- a. A sample of five hospitals is to be randomly selected. The random numbers are 09, 16, 00, 49, 54, 12, and 04. Which hospitals are included in the sample?
  - b. Use a table of random numbers to develop your own sample of five hospitals.
  - c. Using systematic random sampling, every fifth location is selected starting with the second hospital in the list. Which hospitals will be included in the sample?
  - d. Using stratified random sampling, select five hospitals. Four should be medical and surgical hospitals and one should be a specialty hospital. Select an appropriate sample.
3. Listed below are the 35 members of the Metro Toledo Automobile Dealers Association. We would like to estimate the mean revenue from dealer service departments. The members are identified by numbering them 00 through 34.

| ID Number | Dealer                  | ID Number | Dealer                   | ID Number | Dealer                    |
|-----------|-------------------------|-----------|--------------------------|-----------|---------------------------|
| 00        | Dave White Acura        | 12        | Spurgeon Chevrolet Motor | 24        | Lexus of Toledo           |
| 01        | Autofair Nissan         | 13        | Dunn Chevrolet           | 25        | Mathews Ford Oregon Inc.  |
| 02        | Autofair Toyota-Suzuki  | 14        | Don Scott Chevrolet      | 26        | Northtown Chevrolet       |
| 03        | George Ball's Buick GMC | 15        | Dave White Chevrolet Co. | 27        | Quality Ford Sales Inc.   |
| 04        | York Automotive Group   | 16        | Dick Wilson Infiniti     | 28        | Rouen Chrysler Jeep Eagle |
| 05        | Bob Schmidt Chevrolet   | 17        | Doyle Buick              | 29        | Mercedes of Toledo        |
| 06        | Bowling Green Lincoln   | 18        | Franklin Park Lincoln    | 30        | Ed Schmidt Jeep Eagle     |
| 07        | Brondes Ford            | 19        | Genoa Motors             | 31        | Southside Lincoln         |
| 08        | Brown Honda             | 20        | Great Lakes Ford Nissan  | 32        | Valiton Chrysler          |
| 09        | Brown Mazda             | 21        | Grogan Towne Chrysler    | 33        | Vin Divers                |
| 10        | Charlie's Dodge         | 22        | Hatfield Motor Sales     | 34        | Whitman Ford              |
| 11        | Thayer Chevrolet/Toyota | 23        | Kistler Ford Inc.        |           |                           |

- a. We want to select a random sample of five dealers. The random numbers are 05, 20, 59, 21, 31, 28, 49, 38, 66, 08, 29, and 02. Which dealers would be included in the sample?

- b. Use the table of random numbers to select your own sample of five dealers.
  - c. Using systematic random sampling, every seventh dealer is selected starting with the fourth dealer in the list. Which dealers are included in the sample?
4. Listed next are the 27 Nationwide Insurance agents in the El Paso, Texas, metropolitan area. The agents are numbered 00 through 26. We would like to estimate the mean number of years employed with Nationwide.

| ID Number | Agent   | ID Number | Agent  | ID Number | Agent                                 |
|-----------|---|-----------|--|-----------|---------------------------------------|
| 00        | <b>Bly Scott</b> 3332 W Laskey Rd             | 10        | <b>Heini Bernie</b> 7110 W Central Av            | 20        | <b>Schwab Dave</b> 572 W Dussel Dr    |
| 01        | <b>Coyle Mike</b> 5432 W Central Av           | 11        | <b>Hinckley Dave</b><br>14 N Holland Sylvania Rd | 21        | <b>Seibert John H</b> 201 S Main St   |
| 02        | <b>Denker Brett</b> 7445 Airport Hwy          | 12        | <b>Joehlin Bob</b> 3358 Navarre Av               | 22        | <b>Smithers Bob</b> 229 Superior St   |
| 03        | <b>Denker Rollie</b> 7445 Airport Hwy         | 13        | <b>Keisser David</b> 3030 W Sylvania Av          | 23        | <b>Smithers Jerry</b> 229 Superior St |
| 04        | <b>Farley Ron</b> 1837 W Alexis Rd            | 14        | <b>Keisser Keith</b> 5902 Sylvania Av            | 24        | <b>Wright Steve</b> 105 S Third St    |
| 05        | <b>George Mark</b> 7247 W Central Av          | 15        | <b>Lawrence Grant</b> 342 W Dussel Dr            | 25        | <b>Wood Tom</b> 112 Louisiana Av      |
| 06        | <b>Gibellato Carlo</b> 6616 Monroe St         | 16        | <b>Miller Ken</b> 2427 Woodville Rd              | 26        | <b>Yoder Scott</b> 6 Willoughby Av    |
| 07        | <b>Glemser Cathy</b> 5602 Woodville Rd        | 17        | <b>O'Donnell Jim</b> 7247 W Central Av           |           |                                       |
| 08        | <b>Green Mike</b><br>4149 Holland Sylvania Rd | 18        | <b>Priest Harvey</b> 5113 N Summit St            |           |                                       |
| 09        | <b>Harris Ev</b> 2026 Albon Rd                | 19        | <b>Riker Craig</b> 2621 N Reynolds Rd            |           |                                       |

- a. We want to select a random sample of four agents. The random numbers are 02, 59, 51, 25, 14, 29, 77, 69, and 18. Which dealers would be included in the sample?
- b. Use the table of random numbers to select your own sample of four agents.
- c. Using systematic random sampling, every fifth dealer is selected starting with the third dealer in the list. Which dealers are included in the sample?

**LO8-2**

Define sampling error.

## SAMPLING “ERROR”

In the previous section, we discussed sampling methods that are used to select a sample that is an unbiased representation of the population. In each method, the selection of every possible sample of a specified size from a population has a known chance or probability. This is another way to describe an unbiased sampling method.

Samples are used to estimate population characteristics. For example, the mean of a sample is used to estimate the population mean. However, since the sample is a part or portion of the population, it is unlikely that the sample mean would be *exactly equal* to the population mean. Similarly, it is unlikely that the sample standard deviation would be *exactly equal* to the population standard deviation. We can therefore expect a difference between a *sample statistic* and its corresponding *population parameter*. This difference is called **sampling error**.

**SAMPLING ERROR** The difference between a sample statistic and its corresponding population parameter.

The following example/solution clarifies the idea of sampling error.

**EXAMPLE**

Refer to the example/solution on page 213, where we studied the number of rooms rented at the Foxtrot Inn bed and breakfast in Tryon, North Carolina. The population is the number of rooms rented each of the 30 days in June 2017. Find the mean of

the population. Select three random samples of 5 days. Calculate the mean rooms rented for each sample and compare it to the population mean. What is the sampling error in each case?

### SOLUTION

During the month, there were a total of 94 rentals. So the mean number of units rented per night is 3.13. This is the population mean. Hence we designate this value with the Greek letter  $\mu$ .

$$\mu = \frac{\Sigma X}{N} = \frac{0 + 2 + 3 + \cdots + 3}{30} = \frac{94}{30} = 3.13$$

The first random sample of five nights resulted in the following number of rooms rented: 4, 7, 4, 3, and 1. The mean of this sample is 3.80 rooms, which we designate as  $\bar{x}_1$ . The bar over the  $x$  reminds us that it is a sample mean, and the subscript 1 indicates it is the mean of the first sample.

$$\bar{x}_1 = \frac{\Sigma X}{n} = \frac{4 + 7 + 4 + 3 + 1}{5} = \frac{19}{5} = 3.80$$

The sampling error for the first sample is the difference between the population mean (3.13) and the first sample mean (3.80). Hence, the sampling error is  $(\bar{x}_1 - \mu) = 3.80 - 3.13 = 0.67$ . The second random sample of 5 days from the population of all 30 days in June revealed the following number of rooms rented: 3, 3, 2, 3, and 6. The mean of these five values is 3.40, found by

$$\bar{x}_2 = \frac{\Sigma X}{n} = \frac{3 + 3 + 2 + 3 + 6}{5} = 3.40$$

The sampling error is  $(\bar{x}_2 - \mu) = 3.4 - 3.13 = 0.27$ . In the third random sample, the mean was 1.80 and the sampling error was  $-1.33$ .

Each of these differences, 0.67, 0.27, and  $-1.33$ , is the sampling error made in estimating the population mean. Sometimes these errors are positive values, indicating that the sample mean overestimated the population mean; other times they are negative values, indicating the sample mean was less than the population mean.

| June | Rentals | June | Rentals | June | Rentals |  | Sample 1              | Sample 2 | Sample 3 |       |
|------|---------|------|---------|------|---------|--|-----------------------|----------|----------|-------|
| 1    | 0       | 11   | 3       | 21   | 3       |  | 4                     | 3        | 0        |       |
| 2    | 2       | 12   | 4       | 22   | 2       |  | 7                     | 3        | 0        |       |
| 3    | 3       | 13   | 4       | 23   | 3       |  | 4                     | 2        | 3        |       |
| 4    | 2       | 14   | 4       | 24   | 6       |  | 3                     | 3        | 3        |       |
| 5    | 3       | 15   | 7       | 25   | 0       |  | 1                     | 6        | 3        |       |
| 6    | 4       | 16   | 0       | 26   | 4       |  |                       |          |          |       |
| 7    | 2       | 17   | 5       | 27   | 1       |  |                       |          |          |       |
| 8    | 3       | 18   | 3       | 28   | 1       |  |                       |          |          |       |
| 9    | 4       | 19   | 6       | 29   | 3       |  |                       |          |          |       |
| 10   | 7       | 20   | 2       | 30   | 3       |  |                       |          |          |       |
|      |         |      |         |      |         |  | <b>Total</b>          | 19       | 17       | 9     |
|      |         |      |         |      |         |  | <b>Mean</b>           | 3.80     | 3.40     | 1.80  |
|      |         |      |         |      |         |  | <b>Sampling Error</b> | 0.67     | 0.27     | -1.33 |

In this case, where we have a population of 30 values and samples of 5 values, there is a very large number of possible samples—142,506 to be exact! To find this value, use the combination formula (5–9) on page 145. Each of the 142,506 different samples has the same chance of being selected. Each sample may have a different sample mean and therefore a different sampling error. The value of the sampling error is based on the particular one of the 142,506 different possible samples selected. Therefore, the sampling errors are random and occur by chance. If you summed the sampling errors for all 142,506 samples, the result would equal zero. This is true because the sample mean is an *unbiased estimator* of the population mean.

**LO8-3**

Explain the sampling distribution of the sample mean.

## SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

In the previous section, we defined sampling error and presented the results when we compared a sample statistic, such as the sample mean, to the population mean. To put it another way, when we use the sample mean to estimate the population mean, how can we determine how accurate the estimate is? How does:

- A quality-assurance supervisor decide if a machine is filling 20-ounce bottles with 20 ounces of cola based only on a sample of 10 filled bottles?
- [www.FiveThirtyEight.com](http://www.FiveThirtyEight.com) or [www.Gallup.com](http://www.Gallup.com) make accurate statements about the demographics of voters in a presidential race based on relatively small samples from a voting population of nearly 90 million?

To answer these questions, we first develop a *sampling distribution of the sample mean*.

The sample means in the previous example/solution varied from one sample to the next. The mean of the first sample of 5 days was 3.80 rooms, and the second sample mean was 3.40 rooms. The population mean was 3.13 rooms. If we organized the means of all possible samples of 5 days into a probability distribution, the result is called the **sampling distribution of the sample mean**.

**SAMPLING DISTRIBUTION OF THE SAMPLE MEAN** A probability distribution of all possible sample means of a given sample size.

The following example/solution illustrates the construction of a sampling distribution of the sample mean. We have intentionally used a small population to highlight the relationship between the population mean and the various sample means.

**EXAMPLE**

Tartus Industries has seven production employees (considered the population). The hourly earnings of each employee are given in Table 8–2.

**TABLE 8–2** Hourly Earnings of the Production Employees of Tartus Industries

| Employee | Hourly Earnings | Employee | Hourly Earnings |
|----------|-----------------|----------|-----------------|
| Joe      | \$14            | Jan      | \$14            |
| Sam      | 14              | Art      | 16              |
| Sue      | 16              | Ted      | 18              |
| Bob      | 16              |          |                 |

1. What is the population mean?
2. What is the sampling distribution of the sample mean for samples of size 2?
3. What is the mean of the sampling distribution?
4. What observations can be made about the population and the sampling distribution?

**SOLUTION**

1. The population is small, so it is easy to calculate the population mean. It is \$15.43, found by:

$$\mu = \frac{\sum x}{N} = \frac{\$14 + \$14 + \$16 + \$16 + \$14 + \$16 + \$18}{7} = \$15.43$$

We identify the population mean with the Greek letter  $\mu$ . Recall from earlier chapters, Greek letters are used to represent population parameters.

- To arrive at the sampling distribution of the sample mean, we need to select all possible samples of 2 without replacement from the population, then compute the mean of each sample. There are 21 possible samples, found by using formula (5–9) on page 145.

$${}^N C_n = \frac{N!}{n!(N - n)!} = \frac{7!}{2!(7 - 2)!} = 21$$

where  $N = 7$  is the number of items in the population and  $n = 2$  is the number of items in the sample.

**TABLE 8–3** Sample Means for All Possible Samples of 2 Employees

| Sample | Employees | Hourly     |      |      | Sample | Employees | Hourly     |      |      |
|--------|-----------|------------|------|------|--------|-----------|------------|------|------|
|        |           | Earnings   | Sum  | Mean |        |           | Earnings   | Sum  | Mean |
| 1      | Joe, Sam  | \$14, \$14 | \$28 | \$14 | 12     | Sue, Bob  | \$16, \$16 | \$32 | \$16 |
| 2      | Joe, Sue  | 14, 16     | 30   | 15   | 13     | Sue, Jan  | 16, 14     | 30   | 15   |
| 3      | Joe, Bob  | 14, 16     | 30   | 15   | 14     | Sue, Art  | 16, 16     | 32   | 16   |
| 4      | Joe, Jan  | 14, 14     | 28   | 14   | 15     | Sue, Ted  | 16, 18     | 34   | 17   |
| 5      | Joe, Art  | 14, 16     | 30   | 15   | 16     | Bob, Jan  | 16, 14     | 30   | 15   |
| 6      | Joe, Ted  | 14, 18     | 32   | 16   | 17     | Bob, Art  | 16, 16     | 32   | 16   |
| 7      | Sam, Sue  | 14, 16     | 30   | 15   | 18     | Bob, Ted  | 16, 18     | 34   | 17   |
| 8      | Sam, Bob  | 14, 16     | 30   | 15   | 19     | Jan, Art  | 14, 16     | 30   | 15   |
| 9      | Sam, Jan  | 14, 14     | 28   | 14   | 20     | Jan, Ted  | 14, 18     | 32   | 16   |
| 10     | Sam, Art  | 14, 16     | 30   | 15   | 21     | Art, Ted  | 16, 18     | 34   | 17   |
| 11     | Sam, Ted  | 14, 18     | 32   | 16   |        |           |            |      |      |

The 21 sample means from all possible samples of 2 that can be drawn from the population of 7 employees are shown in Table 8–3. These 21 sample means are used to construct a probability distribution. This is called the sampling distribution of the sample mean, and it is summarized in Table 8–4.

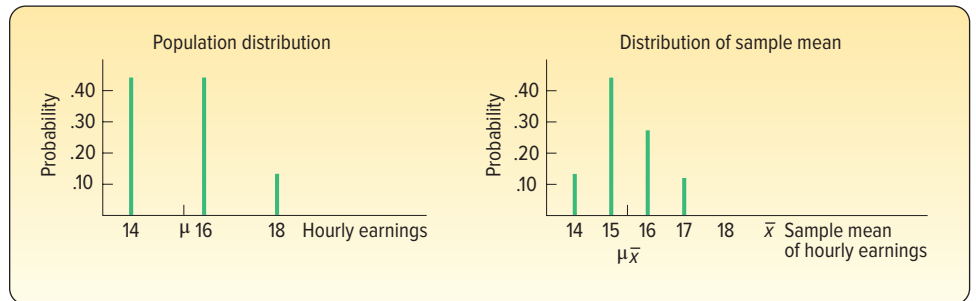
**TABLE 8–4** Sampling Distribution of the Sample Mean for  $n = 2$

| Sample Mean | Number of Means | Probability   |
|-------------|-----------------|---------------|
| \$14        | 3               | .1429         |
| 15          | 9               | .4285         |
| 16          | 6               | .2857         |
| 17          | 3               | .1429         |
|             | <u>21</u>       | <u>1.0000</u> |

- Using the data in Table 8–3, the mean of the sampling distribution of the sample mean is obtained by summing the various sample means and dividing the sum by the number of samples. The mean of all the sample means is usually written  $\mu_{\bar{x}}$ . The  $\mu$  reminds us that it is a population value because we have considered all possible samples of two employees from the population of seven employees. The subscript  $\bar{x}$  indicates that it is the sampling distribution of the sample mean.

$$\begin{aligned} \mu_{\bar{x}} &= \frac{\text{Sum of all sample means}}{\text{Total number of samples}} = \frac{\$14 + \$15 + \$15 + \dots + \$16 + \$17}{21} \\ &= \frac{\$324}{21} = \$15.43 \end{aligned}$$

4. Refer to Chart 8–2. It shows the population distribution based on the data in Table 8–2 and the distribution of the sample mean based on the data in Table 8–4. These observations can be made:
  - a. The mean of the distribution of the sample mean (\$15.43) is equal to the mean of the population:  $\mu = \mu_{\bar{x}}$ .
  - b. The spread in the distribution of the sample mean is less than the spread in the population values. The sample means range from \$14 to \$17, while the population values vary from \$14 up to \$18. If we continue to increase the sample size, the spread of the distribution of the sample mean becomes smaller.
  - c. The shape of the sampling distribution of the sample mean and the shape of the frequency distribution of the population values are different. The distribution of the sample mean tends to be more bell-shaped and to approximate the normal probability distribution.



**CHART 8–2** Distributions of Population Values and Sample Means

In summary, we took all possible random samples from a population and for each sample calculated a sample statistic (the mean amount earned). This example illustrates important relationships between the population distribution and the sampling distribution of the sample mean:

1. The mean of the sample means is exactly equal to the population mean.
2. The dispersion of the sampling distribution of the sample mean is narrower than the population distribution.
3. The sampling distribution of the sample mean tends to become bell-shaped and to approximate a normal probability distribution.

Given a bell-shaped or normal probability distribution, we will be able to apply concepts from Chapter 7 to determine the probability of selecting a sample with a specified sample mean. In the next section, we will show the importance of sample size as it relates to the sampling distribution of the sample mean.

### SELF-REVIEW 8–3

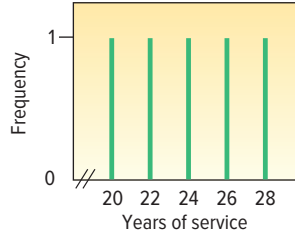


The years of service of the five executives employed by Standard Chemicals are:

| Name       | Years |
|------------|-------|
| Mr. Snow   | 20    |
| Ms. Tolson | 22    |
| Mr. Kraft  | 26    |
| Ms. Irwin  | 24    |
| Mr. Jones  | 28    |



- (a) Using the combination formula, how many samples of size 2 are possible?
- (b) List all possible samples of two executives from the population and compute their means.
- (c) Organize the means into a sampling distribution.
- (d) Compare the population mean and the mean of the sample means.
- (e) Compare the dispersion in the population with that in the distribution of the sample mean.
- (f) A chart portraying the population values follows. Is the distribution of population values normally distributed (bell-shaped)?



- (g) Is the distribution of the sample mean computed in part (c) starting to show some tendency toward a normal distribution?

## EXERCISES

- 5. A population consists of the following four values: 12, 12, 14, and 16.
  - a. List all samples of size 2, and compute the mean of each sample.
  - b. Compute the mean of the distribution of the sample mean and the population mean. Compare the two values.
  - c. Compare the dispersion in the population with that of the sample mean.
- 6. A population consists of the following five values: 2, 2, 4, 4, and 8.
  - a. List all samples of size 2, and compute the mean of each sample.
  - b. Compute the mean of the distribution of sample means and the population mean. Compare the two values.
  - c. Compare the dispersion in the population with that of the sample means.
- 7. A population consists of the following five values: 12, 12, 14, 15, and 20.
  - a. List all samples of size 3, and compute the mean of each sample.
  - b. Compute the mean of the distribution of sample means and the population mean. Compare the two values.
  - c. Compare the dispersion in the population with that of the sample means.
- 8. A population consists of the following five values: 0, 0, 1, 3, and 6.
  - a. List all samples of size 3, and compute the mean of each sample.
  - b. Compute the mean of the distribution of sample means and the population mean. Compare the two values.
  - c. Compare the dispersion in the population with that of the sample means.
- 9. In the law firm Tybo and Associates, there are six partners. Listed is the number of cases each partner actually tried in court last month.

| Partner   | Number of Cases |
|-----------|-----------------|
| Ruud      | 3               |
| Wu        | 6               |
| Sass      | 3               |
| Flores    | 3               |
| Wilhelms  | 0               |
| Schueller | 1               |

- a. How many different samples of size 3 are possible?
- b. List all possible samples of size 3, and compute the mean number of cases in each sample.
- c. Compare the mean of the distribution of sample means to the population mean.
- d. On a chart similar to Chart 8–2, compare the dispersion in the population with that of the sample means.

10. There are five sales associates at Mid-Motors Ford. The five associates and the number of cars they sold last week are:

| Sales Associate | Cars Sold |
|-----------------|-----------|
| Peter Hankish   | 8         |
| Connie Stallter | 6         |
| Juan Lopez      | 4         |
| Ted Barnes      | 10        |
| Peggy Chu       | 6         |

- How many different samples of size 2 are possible?
- List all possible samples of size 2, and compute the mean of each sample.
- Compare the mean of the sampling distribution of the sample mean with that of the population.
- On a chart similar to Chart 8–2, compare the dispersion in sample means with that of the population.

#### LO8-4

Recite the central limit theorem and define the mean and standard error of the sampling distribution of the sample mean.

## THE CENTRAL LIMIT THEOREM

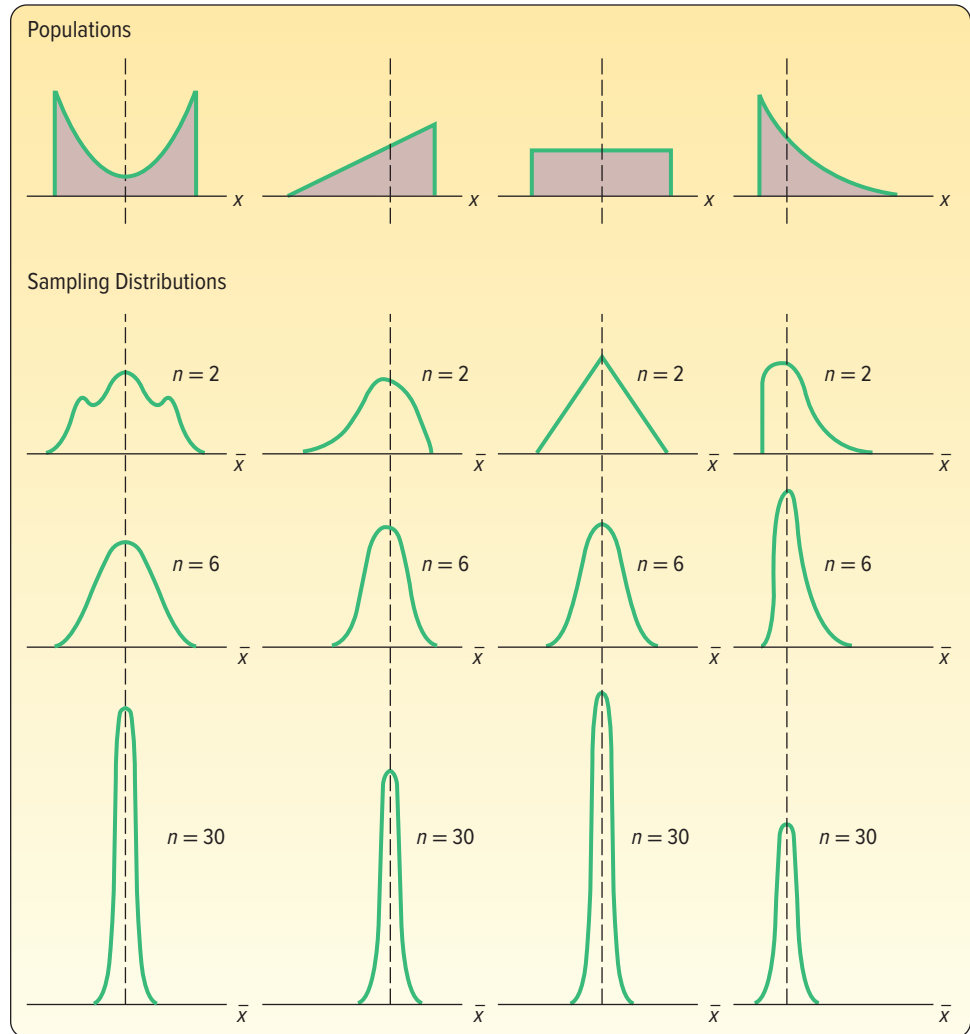
In this section, we examine the **central limit theorem**. Its application to the sampling distribution of the sample mean, introduced in the previous section, allows us to use the normal probability distribution to create confidence intervals for the population mean (described in Chapter 9) and perform tests of hypothesis (described in Chapter 10). The central limit theorem states that, for large random samples, the shape of the sampling distribution of the sample mean is close to the normal probability distribution. The approximation is more accurate for large samples than for small samples. This is one of the most useful conclusions in statistics. We can reason about the distribution of the sample mean with absolutely no information about the shape of the population distribution from which the sample is taken. In other words, the central limit theorem is true for all population distributions.

**CENTRAL LIMIT THEOREM** If all samples of a particular size are selected from any population, the sampling distribution of the sample mean is approximately a normal distribution. This approximation improves with larger samples.

To further illustrate the central limit theorem, if the population follows a normal probability distribution, then for any sample size, the sampling distribution of the sample mean will also be normal. If the population distribution is symmetrical (but not normal), you will see the normal shape of the distribution of the sample mean emerge with samples as small as 10. On the other hand, if you start with a distribution that is skewed or has thick tails, it may require samples of 30 or more to observe the normality feature. This concept is summarized in Chart 8–3 for various population shapes. Observe the convergence to a normal distribution regardless of the shape of the population distribution. The following example/solution will illustrate this concept.

### EXAMPLE

Ed Spence began his sprocket business 20 years ago. The business has grown over the years and now employs 40 people. Spence Sprockets Inc. faces some major decisions regarding health care for these employees. Before making a final



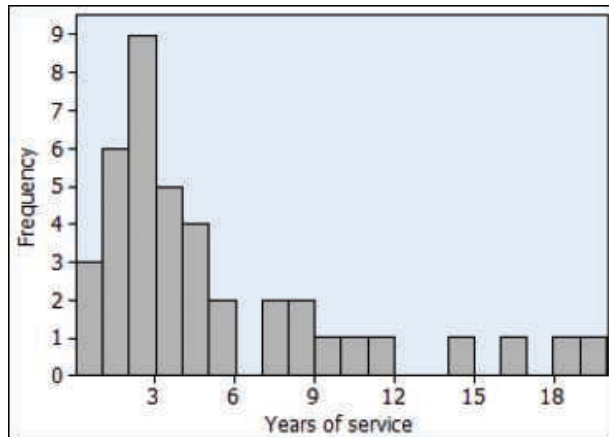
**CHART 8-3** Results of the Central Limit Theorem for Several Populations

decision on what health care plan to purchase, Ed decides to form a committee of five representative employees. The committee will be asked to study the health care issue carefully and make a recommendation as to what plan best fits the employees' needs. Ed feels the views of newer employees toward health care may differ from those of more experienced employees. If Ed randomly selects this committee, what can he expect in terms of the mean years with Spence Sprockets for those on the committee? How does the shape of the distribution of years of service of all employees (the population) compare with the shape of the sampling distribution of the mean? The years of service (rounded to the nearest year) of the 40 employees currently on the Spence Sprockets Inc. payroll are as follows.

|    |   |    |   |   |   |   |    |   |    |
|----|---|----|---|---|---|---|----|---|----|
| 11 | 4 | 18 | 2 | 1 | 2 | 0 | 2  | 2 | 4  |
| 3  | 4 | 1  | 2 | 2 | 3 | 3 | 19 | 8 | 3  |
| 7  | 1 | 0  | 2 | 7 | 0 | 4 | 5  | 1 | 14 |
| 16 | 8 | 9  | 1 | 1 | 2 | 5 | 10 | 2 | 3  |

### SOLUTION

Chart 8–4 shows a histogram for the frequency distribution of the years of service for the population of 40 current employees. This distribution is positively skewed. Why? Because the business has grown in recent years, the distribution shows that 29 of the 40 employees have been with the company less than 6 years. Also, there are 11 employees who have worked at Spence Sprockets for more than 6 years. In particular, four employees have been with the company 12 years or more (count the frequencies above 12). So there is a long tail in the distribution of service years to the right, that is, the distribution is positively skewed.



**CHART 8–4** Years of Service for Spence Sprockets Inc. Employees

Let's consider the first of Ed Spence's problems. He would like to form a committee of five employees to look into the health care question and suggest what type of health care coverage would be most appropriate for the majority of workers. How should he select the committee? If he selects the committee randomly, what might he expect in terms of mean years of service for those on the committee?

To begin, Ed writes the years of service for each of the 40 employees on pieces of paper and puts them into an old baseball hat. Next, he shuffles the pieces of paper and randomly selects five slips of paper. The years of service for these five employees are 1, 9, 0, 19, and 14 years. Thus, the mean years of service for these five sampled employees is 8.60 years. How does that compare with the population mean? At this point, Ed does not know the population mean, but the number of employees in the population is only 40, so he decides to calculate the mean years of service for *all* his employees. It is 4.8 years, found by adding the years of service for *all* the employees and dividing the total by 40.

$$\mu = \frac{11 + 4 + 18 + \cdots + 2 + 3}{40} = 4.80$$

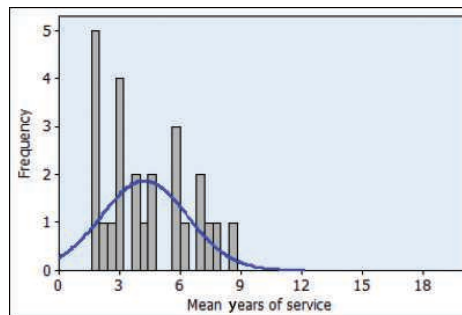
The difference between a sample mean ( $\bar{x}$ ) and the population mean ( $\mu$ ) is called **sampling error**. In other words, the difference of 3.80 years between the sample mean of 8.60 and the population mean of 4.80 is the sampling error. It is due to chance. Thus, if Ed selected these five employees to constitute the committee, their mean years of service would be larger than the population mean.

What would happen if Ed put the five pieces of paper back into the baseball hat and selected another sample? Would you expect the mean of this second sample to be exactly the same as the previous one? Suppose he selects another sample of five employees and finds the years of service in this sample to be 7, 4, 4, 1, and 3. This sample mean is 3.80 years. The result of selecting 25 samples of five employees

**TABLE 8-5** Twenty-Five Random Samples of Five Employees

| Sample Data |       |       |       |       |       |     |      |
|-------------|-------|-------|-------|-------|-------|-----|------|
| Sample      | Obs 1 | Obs 2 | Obs 3 | Obs 4 | Obs 5 | Sum | Mean |
| A           | 1     | 9     | 0     | 19    | 14    | 43  | 8.6  |
| B           | 7     | 4     | 4     | 1     | 3     | 19  | 3.8  |
| C           | 8     | 19    | 8     | 2     | 1     | 38  | 7.6  |
| D           | 4     | 18    | 2     | 0     | 11    | 35  | 7.0  |
| E           | 4     | 2     | 4     | 7     | 18    | 35  | 7.0  |
| F           | 1     | 2     | 0     | 3     | 2     | 8   | 1.6  |
| G           | 2     | 3     | 2     | 0     | 2     | 9   | 1.8  |
| H           | 11    | 2     | 9     | 2     | 4     | 28  | 5.6  |
| I           | 9     | 0     | 4     | 2     | 7     | 22  | 4.4  |
| J           | 1     | 1     | 1     | 11    | 1     | 15  | 3.0  |
| K           | 2     | 0     | 0     | 10    | 2     | 14  | 2.8  |
| L           | 0     | 2     | 3     | 2     | 16    | 23  | 4.6  |
| M           | 2     | 3     | 1     | 1     | 1     | 8   | 1.6  |
| N           | 3     | 7     | 3     | 4     | 3     | 20  | 4.0  |
| O           | 1     | 2     | 3     | 1     | 4     | 11  | 2.2  |
| P           | 19    | 0     | 1     | 3     | 8     | 31  | 6.2  |
| Q           | 5     | 1     | 7     | 14    | 9     | 36  | 7.2  |
| R           | 5     | 4     | 2     | 3     | 4     | 18  | 3.6  |
| S           | 14    | 5     | 2     | 2     | 5     | 28  | 5.6  |
| T           | 2     | 1     | 1     | 4     | 7     | 15  | 3.0  |
| U           | 3     | 7     | 1     | 2     | 1     | 14  | 2.8  |
| V           | 0     | 1     | 5     | 1     | 2     | 9   | 1.8  |
| W           | 0     | 3     | 19    | 4     | 2     | 28  | 5.6  |
| X           | 4     | 2     | 3     | 4     | 0     | 13  | 2.6  |
| Y           | 1     | 1     | 2     | 3     | 2     | 9   | 1.8  |

and computing the mean for each sample is shown in Table 8-5 and Chart 8-5. There are actually 658,008 possible samples of 5 from the population of 40 employees, found by the combination formula (5-9) for 40 things taken 5 at a time. Notice the difference in the shape of the population and the distribution of these sample means. The population of the years of service for employees (Chart 8-4) is positively skewed, but the distribution of these 25 sample means does not reflect the same positive skew. There is also a difference in the range of the sample means versus the range of the population. The population age varies from 0 to 19, so the population's range is 19 years. The sample means vary from 1.6 to 8.6 years, so the range of the sample means is 7 years. The dispersion of the sample means is less than the dispersion of values in the population.

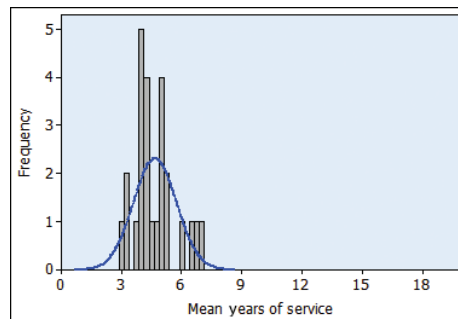
**CHART 8-5** Histogram of Mean Years of Service for 25 Samples of Five Employees

**TABLE 8-6** Twenty-Five Random Samples of 20 Employees

| Sample Data |       |       |       |   |        |        |     |      |
|-------------|-------|-------|-------|---|--------|--------|-----|------|
| Sample      | Obs 1 | Obs 2 | Obs 3 | – | Obs 19 | Obs 20 | Sum | Mean |
| A           | 3     | 8     | 3     | – | 4      | 16     | 79  | 3.95 |
| B           | 2     | 3     | 8     | – | 3      | 1      | 65  | 3.25 |
| C           | 14    | 5     | 0     | – | 19     | 8      | 119 | 5.95 |
| D           | 9     | 2     | 1     | – | 1      | 3      | 87  | 4.35 |
| E           | 18    | 1     | 2     | – | 3      | 14     | 107 | 5.35 |
| F           | 10    | 4     | 4     | – | 2      | 1      | 80  | 4.00 |
| G           | 5     | 7     | 11    | – | 2      | 4      | 131 | 6.55 |
| H           | 3     | 0     | 2     | – | 16     | 5      | 85  | 4.25 |
| I           | 0     | 0     | 18    | – | 2      | 3      | 80  | 4.00 |
| J           | 2     | 7     | 2     | – | 3      | 2      | 81  | 4.05 |
| K           | 7     | 4     | 5     | – | 1      | 2      | 84  | 4.20 |
| L           | 0     | 3     | 10    | – | 0      | 4      | 81  | 4.05 |
| M           | 4     | 1     | 2     | – | 1      | 2      | 88  | 4.40 |
| N           | 3     | 16    | 1     | – | 11     | 1      | 95  | 4.75 |
| O           | 2     | 19    | 2     | – | 2      | 2      | 102 | 5.10 |
| P           | 2     | 18    | 16    | – | 4      | 3      | 100 | 5.00 |
| Q           | 3     | 2     | 3     | – | 3      | 1      | 102 | 5.10 |
| R           | 2     | 3     | 1     | – | 0      | 2      | 73  | 3.65 |
| S           | 2     | 14    | 19    | – | 0      | 7      | 142 | 7.10 |
| T           | 0     | 1     | 3     | – | 2      | 0      | 61  | 3.05 |
| U           | 1     | 0     | 1     | – | 9      | 3      | 65  | 3.25 |
| V           | 1     | 9     | 4     | – | 2      | 11     | 137 | 6.85 |
| W           | 8     | 1     | 9     | – | 8      | 7      | 107 | 5.35 |
| X           | 4     | 2     | 0     | – | 2      | 5      | 86  | 4.30 |
| Y           | 1     | 2     | 1     | – | 1      | 18     | 101 | 5.05 |

Now let's change the example by increasing the size of each sample from 5 employees to 20. Table 8-6 reports the result of selecting 25 samples of 20 employees each and computing their sample means. These sample means are shown graphically in Chart 8-6. Compare the shape of this distribution to the population (Chart 8-4) and to the distribution of sample means where the sample is  $n = 5$  (Chart 8-5). You should observe two important features:

1. The shape of the distribution of the sample mean is different from that of the population. In Chart 8-4, the distribution of all employees is positively skewed. However, as we select random samples from this population, the shape of the distribution of the sample mean changes. As we increase the size of the sample, the distribution of the sample mean approaches the normal probability distribution. This illustrates the central limit theorem.



**CHART 8-6** Histogram of Mean Years of Service for 25 Samples of 20 Employees

2. There is less dispersion in the sampling distribution of the sample mean than in the population distribution. In the population, the years of service varied from 0 to 19 years. When we selected samples of 5, the sample means varied from 1.6 to 8.6 years, and when we selected samples of 20, the means varied from 3.05 to 7.10 years.

We can also compare the mean of the sample means to the population mean. The mean of the 25 samples of 20 employees reported in Table 8–6 is 4.676 years.

$$\mu_{\bar{x}} = \frac{3.95 + 3.25 + \cdots + 4.30 + 5.05}{25} = 4.676$$

We use the symbol  $\mu_{\bar{x}}$  to identify the mean of the distribution of the sample mean. The subscript reminds us that the distribution is of the sample mean. It is read “mu sub x bar.” We observe the mean of the sample means, 4.676 years, is very close to the population mean of 4.80.

What should we conclude from this example? The central limit theorem indicates that, regardless of the shape of the population distribution, the sampling distribution of the sample mean will move toward a normal probability distribution. The larger the number of observations sampled or selected, the stronger the convergence. The Spence Sprockets Inc. example shows how the central limit theorem works. We began with a positively skewed population (Chart 8–4). Next, we selected 25 random samples of 5 observations, computed the mean of each sample, and finally organized these 25 sample means into a histogram (Chart 8–5). We observe that the shape of the sampling distribution of the sample mean is very different from that of the population. The population distribution is positively skewed compared to the nearly normal shape of the sampling distribution of the sample mean.

To further illustrate the effects of the central limit theorem, we increased the number of observations in each sample from 5 to 20. We selected 25 samples of 20 observations each and calculated the mean of each sample. Finally, we organized these sample means into a histogram (Chart 8–6). The shape of the histogram in Chart 8–6 is clearly moving toward the normal probability distribution.

If you go back to Chapter 6, where several binomial distributions with a “success” proportion of .10 are shown in Chart 6–3 on page 169, you can see yet another demonstration of the central limit theorem. Observe as  $n$  increases from 7 through 12 and 20 up to 40 that the profile of the probability distributions moves closer and closer to a normal probability distribution. Chart 8–6 also shows the convergence to normality as  $n$  increases. This again reinforces the fact that, as more observations are sampled from any population distribution, the shape of the sampling distribution of the sample mean will get closer and closer to a normal distribution.

The **central limit theorem**, defined on page 225, does not say anything about the dispersion of the sampling distribution of the sample mean or about the comparison of the mean of the sampling distribution of the sample mean to the mean of the population. However, in our Spence Sprockets example, we did observe that there was less dispersion in the distribution of the sample mean than in the population distribution by noting the difference in the range in the population and the range of the sample means. We observe that the mean of the sample means is close to the mean of the population. It can be demonstrated that the mean of the sampling distribution is exactly equal to the population mean (i.e.,  $\mu_{\bar{x}} = \mu$ ), and if the standard deviation in the population is  $\sigma$ , the standard deviation of the sample means is  $\sigma/\sqrt{n}$  where  $n$  is the number of observations in each sample. We refer to  $\sigma/\sqrt{n}$  as the **standard error of the mean**. Its longer name is actually the *standard deviation of the sampling distribution of the sample mean*.

**STANDARD ERROR OF THE MEAN**  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$  **(8-1)**

In this section, we also came to other important conclusions.

1. The mean of the distribution of sample means will be *exactly* equal to the population mean if we are able to select all possible samples of the same size from a given population. That is:

$$\mu = \mu_{\bar{x}}$$

Even if we do not select all samples, we can expect the mean of the distribution of sample means to be close to the population mean.

2. There will be less dispersion in the sampling distribution of the sample mean than in the population. If the standard deviation of the population is  $\sigma$ , the standard deviation of the distribution of sample means is  $\sigma/\sqrt{n}$ . Note that when we increase the size of the sample, the standard error of the mean decreases.

### SELF-REVIEW 8-4



Refer to the Spence Sprockets Inc. data on page 226. Select 10 random samples of five employees each. Use the methods described earlier in the chapter and the Table of Random Numbers (Appendix B.4) to find the employees to include in the sample. Compute the mean of each sample and plot the sample means on a chart similar to Chart 8-4. What is the mean of your 10 sample means?

### EXERCISES

11. **FILE** Appendix B.4 is a table of random numbers that are uniformly distributed. Hence, each digit from 0 to 9 has the same likelihood of occurrence.
  - a. Draw a graph showing the population distribution of random numbers. What is the population mean?
  - b. Following are the first 10 rows of five digits from the table of random numbers in Appendix B.4. Assume that these are 10 random samples of five values each. Determine the mean of each sample and plot the means on a chart similar to Chart 8-4. Compare the mean of the sampling distribution of the sample mean with the population mean.

|   |   |   |   |   |
|---|---|---|---|---|
| 0 | 2 | 7 | 1 | 1 |
| 9 | 4 | 8 | 7 | 3 |
| 5 | 4 | 9 | 2 | 1 |
| 7 | 7 | 6 | 4 | 0 |
| 6 | 1 | 5 | 4 | 5 |
| 1 | 7 | 1 | 4 | 7 |
| 1 | 3 | 7 | 4 | 8 |
| 8 | 7 | 4 | 5 | 5 |
| 0 | 8 | 9 | 9 | 9 |
| 7 | 8 | 8 | 0 | 4 |

12. **FILE** Scrapper Elevator Company has 20 sales representatives who sell its product throughout the United States and Canada. The number of units sold last month by each representative is listed below. Assume these sales figures to be the population values.

2 3 2 3 3 4 2 4 3 2 2 7 3 4 5 3 3 3 3 5



- a. Draw a graph showing the population distribution.
  - b. Compute the mean of the population.
  - c. Select five random samples of 5 each. Compute the mean of each sample. Use the methods described in this chapter and Appendix B.4 to determine the items to be included in the sample.
  - d. Compare the mean of the sampling distribution of the sample mean to the population mean. Would you expect the two values to be about the same?
  - e. Draw a histogram of the sample means. Do you notice a difference in the shape of the distribution of sample means compared to the shape of the population distribution?
13. Consider all of the coins (pennies, nickels, quarters, etc.) in your pocket or purse as a population. Make a frequency table beginning with the current year and counting backward to record the ages (in years) of the coins. For example, if the current year is 2017, then a coin with 2015 stamped on it is 2 years old.
    - a. Draw a histogram or other graph showing the population distribution.
    - b. Randomly select five coins and record the mean age of the sampled coins. Repeat this sampling process 20 times. Now draw a histogram or other graph showing the distribution of the sample means.
    - c. Compare the shapes of the two histograms.
  14. Consider the digits in the phone numbers on a randomly selected page of your local phone book a population. Make a frequency table of the final digit of 30 randomly selected phone numbers. For example, if a phone number is 555-9704, record a 4.
    - a. Draw a histogram or other graph of this population distribution. Using the uniform distribution, compute the population mean and the population standard deviation.
    - b. Also record the sample mean of the final four digits (9704 would lead to a mean of 5). Now draw a histogram or other graph showing the distribution of the sample means.
    - c. Compare the shapes of the two histograms.

**LO8-5**

Apply the central limit theorem to calculate probabilities.

## USING THE SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

The previous discussion is important because most business decisions are made on the basis of sample information. Here are some examples.

1. Arm & Hammer Company wants to ensure that its laundry detergent actually contains 100 fluid ounces, as indicated on the label. Historical summaries from the filling process indicate the mean amount per container is 100 fluid ounces and the standard deviation is 2 fluid ounces. At 10 a.m., a quality technician measures 40 containers and finds the mean amount per container is 99.8 fluid ounces. Should the technician shut down the filling operation?
2. A. C. Nielsen Company provides information to organizations advertising on television. Prior research indicates that adult Americans watch an average of 6.0 hours per day of television. The standard deviation is 1.5 hours. What is the probability that we could randomly select a sample of 50 adults and find that they watch an average of 6.5 hours or more of television per day?
3. Houghton Elevator Company wishes to develop specifications for the number of people who can ride in a new oversized elevator. Suppose the mean weight of an adult is 160 pounds and the standard deviation is 15 pounds. However, the distribution of weights does not follow the normal probability distribution. It is positively skewed. For a sample of 30 adults, what is the likelihood that their mean weight is 170 pounds or more?

We can answer the questions in each of these situations using the ideas discussed in the previous section. In each case, we have a population with information about its mean and standard deviation. Using this information and sample size, we can determine the distribution of sample means and compute the probability that a sample mean will fall within a certain range. The sampling distribution will be normally distributed under two conditions:

1. When the samples are taken from populations known to follow the normal distribution. In this case, the size of the sample is not a factor.
2. When the shape of the population distribution is not known, sample size is important. In general, the sampling distribution will be normally distributed as the sample size approaches infinity. In practice, a sampling distribution will be close to a normal distribution with samples of at least 30 observations.

We use formula (7–5) from the previous chapter to convert any normal distribution to the standard normal distribution. Using formula (7–5) to compute  $z$  values, we can use the standard normal table, Appendix B.3, to find the probability that an observation is within a specific range. The formula for finding a  $z$  value is:

$$z = \frac{x - \mu}{\sigma}$$

In this formula,  $x$  is the value of the random variable,  $\mu$  is the population mean, and  $\sigma$  is the population standard deviation.

However, when we sample from populations, we are interested in the distribution of  $\bar{X}$ , the sample mean, instead of  $X$ , the value of one observation. That is the first change we make in formula (7–5). The second is that we use the standard error of the mean of  $n$  observations instead of the population standard deviation. That is, we use  $\sigma/\sqrt{n}$  in the denominator rather than  $\sigma$ . Therefore, to find the likelihood of a sample mean within a specified range, we first use the following formula to find the corresponding  $z$  value. Then we use Appendix B.3 or statistical software to determine the probability.

**FINDING THE  $z$  VALUE OF  $\bar{x}$  WHEN THE POPULATION STANDARD DEVIATION IS KNOWN**

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad (8-2)$$

The following example/solution will show the application.

### EXAMPLE

The quality assurance department for Cola Inc. maintains records regarding the amount of cola in its jumbo bottle. The actual amount of cola in each bottle is critical but varies a small amount from one bottle to the next. Cola Inc. does not wish to underfill the bottles because it will have a problem with truth in labeling. On the other hand, it cannot overfill each bottle because it would be giving cola away, hence reducing its profits. Records maintained by the quality assurance department indicate that the amount of cola follows the normal probability distribution. The mean amount per bottle is 31.2 ounces and the population standard deviation is 0.4 ounce. At 8 a.m. today the quality technician randomly selected 16 bottles from the filling line. The mean amount of cola contained in the bottles is 31.38 ounces. Is this an unlikely result? Is it likely the process is putting too much soda in the bottles? To put it another way, is the sampling error of 0.18 ounce unusual?

### SOLUTION

We use the results of the previous section to find the likelihood that we could select a sample of 16 ( $n$ ) bottles from a normal population with a mean of 31.2 ( $\mu$ ) ounces

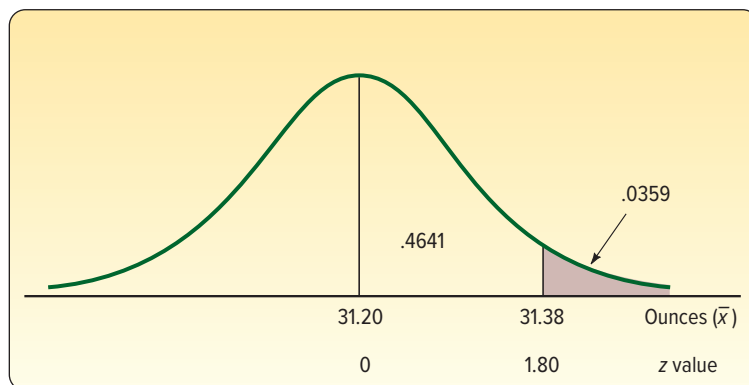
and a population standard deviation of 0.4 ( $\sigma$ ) ounce and find the sample mean to be 31.38 ( $\bar{x}$ ) or more. We use formula (8-2) to find the value of  $z$ .

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{31.38 - 31.20}{0.4/\sqrt{16}} = 1.80$$

The numerator of this equation,  $\bar{x} - \mu = 31.38 - 31.20 = .18$ , is the sampling error. The denominator,  $\sigma/\sqrt{n} = 0.4/\sqrt{16} = 0.1$ , is the standard error of the sampling distribution of the sample mean. So the  $z$  values express the sampling error in standard units—in other words, the standard error.

Next, we compute the likelihood of a  $z$  value greater than 1.80. In Appendix B.3, locate the probability corresponding to a  $z$  value of 1.80. It is .4641. The likelihood of a  $z$  value greater than 1.80 is .0359, found by  $.5000 - .4641$ .

What do we conclude? It is unlikely, less than a 4% chance, we could select a sample of 16 observations from a normal population with a mean of 31.2 ounces and a population standard deviation of 0.4 ounce and find the sample mean equal to or greater than 31.38 ounces. We conclude the process is putting too much cola in the bottles. The quality technician should see the production supervisor about reducing the amount of soda in each bottle. This information is summarized in Chart 8-7.



**CHART 8-7** Sampling Distribution of the Mean Amount of Cola in a Jumbo Bottle

## SELF-REVIEW 8-5



Refer to the Cola Inc. information. Suppose the quality technician selected a sample of 16 jumbo bottles that averaged 31.08 ounces. What can you conclude about the filling process?

## EXERCISES

15. A normal population has a mean of 60 and a standard deviation of 12. You select a random sample of 9. Compute the probability the sample mean is:
  - a. Greater than 63.
  - b. Less than 56.
  - c. Between 56 and 63.
16. A normal population has a mean of 75 and a standard deviation of 5. You select a sample of 40. Compute the probability the sample mean is:
  - a. Less than 74.
  - b. Between 74 and 76.
  - c. Between 76 and 77.
  - d. Greater than 77.

17. In a certain section of Southern California, the distribution of monthly rent for a one-bedroom apartment has a mean of \$2,200 and a standard deviation of \$250. The distribution of the monthly rent does not follow the normal distribution. In fact, it is positively skewed. What is the probability of selecting a sample of 50 one-bedroom apartments and finding the mean to be at least \$1,950 per month?
18. According to an IRS study, it takes a mean of 330 minutes for taxpayers to prepare, copy, and electronically file a 1040 tax form. This distribution of times follows the normal distribution and the standard deviation is 80 minutes. A consumer watchdog agency selects a random sample of 40 taxpayers.
  - a. What is the standard error of the mean in this example?
  - b. What is the likelihood the sample mean is greater than 320 minutes?
  - c. What is the likelihood the sample mean is between 320 and 350 minutes?
  - d. What is the likelihood the sample mean is greater than 350 minutes?

## CHAPTER SUMMARY

- I. There are many reasons for sampling a population.
  - A. The results of a sample may adequately estimate the value of the population parameter, thus saving time and money.
  - B. It may be too time-consuming to contact all members of the population.
  - C. It may be impossible to check or locate all members of the population.
  - D. The cost of studying all the items in the population may be prohibitive.
  - E. Often, testing destroys the sampled item and it cannot be returned to the population.
- II. In an unbiased or probability sample, all members of the population have a chance of being selected for the sample. There are several probability sampling methods.
  - A. In a simple random sample, all members of the population have the same chance of being selected for the sample.
  - B. In a systematic sample, a random starting point is selected, and then every  $k$ th item thereafter is selected for the sample.
  - C. In a stratified sample, the population is divided into several groups, called strata, and then a random sample is selected from each stratum.
  - D. In cluster sampling, the population is divided into primary units, then samples are drawn from the primary units.
- III. The sampling error is the difference between a population parameter and a sample statistic.
- IV. The sampling distribution of the sample mean is a probability distribution of all possible sample means of the same sample size.
  - A. For a given sample size, the mean of all possible sample means selected from a population is equal to the population mean.
  - B. There is less variation in the distribution of the sample mean than in the population distribution.
  - C. The standard error of the mean measures the variation in the sampling distribution of the sample mean. The standard error is found by:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (8-1)$$

- D. If the population follows a normal distribution, the sampling distribution of the sample mean will also follow the normal distribution for samples of any size. If the population is not normally distributed, the sampling distribution of the sample mean will approach a normal distribution when the sample size is at least 30. Assume the population standard deviation is known. To determine the probability that a sample mean falls in a particular region, use the following formula.

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad (8-2)$$

## PRONUNCIATION KEY

| SYMBOL             | MEANING  | PRONUNCIATION          |
|--------------------|--|------------------------|
| $\mu_{\bar{x}}$    | Mean of the sampling distribution of the sample mean | <i>mu sub x bar</i>    |
| $\sigma_{\bar{x}}$ | Population standard error of the sample mean         | <i>sigma sub x bar</i> |

## CHAPTER EXERCISES

19. The 25 retail stores located in the North Towne Square Mall numbered 00 through 24 are:

|                             |                          |                       |
|-----------------------------|--------------------------|-----------------------|
| 00 Elder-Beerman            | 09 Lion Store            | 18 County Seat        |
| 01 Sears                    | 10 Bootleggers           | 19 Kid Mart           |
| 02 Deb Shop                 | 11 Formal Man            | 20 Lerner             |
| 03 Frederick's of Hollywood | 12 Leather Ltd.          | 21 Coach House Gifts  |
| 04 Petries                  | 13 Barnes & Noble        | 22 Spencer Gifts      |
| 05 Easy Dreams              | 14 Pat's Hallmark        | 23 CPI Photo Finish   |
| 06 Summit Stationers        | 15 Things Remembered     | 24 Regis Hairstylists |
| 07 E. B. Brown Opticians    | 16 Pearle Vision Express |                       |
| 08 Kay-Bee Toy & Hobby      | 17 Dollar Tree           |                       |

- a. If the following random numbers are selected, which retail stores should be contacted for a survey? 11, 65, 86, 62, 06, 10, 12, 77, and 04
- b. Select a random sample of four retail stores. Use Appendix B.4.
- c. A systematic sampling procedure will be used. The first store will be selected and then every third store. Which stores will be in the sample?
20. The Medical Assurance Company is investigating the cost of a routine office visit to family practice physicians in the Rochester, New York, area. The following is a list of 39 family practice physicians in the region. Physicians are to be randomly selected and contacted regarding their charges. The 39 physicians have been coded from 00 to 38. Also noted is whether they are in practice by themselves (S), have a partner (P), or are in a group practice (G).

| Number | Physician                | Type of Practice | Number | Physician                 | Type of Practice |
|--------|--------------------------|------------------|--------|---------------------------|------------------|
| 00     | R. E. Scherbarth, M.D.   | S                | 20     | Gregory Yost, M.D.        | P                |
| 01     | Crystal R. Goveia, M.D.  | P                | 21     | J. Christian Zona, M.D.   | P                |
| 02     | Mark D. Hillard, M.D.    | P                | 22     | Larry Johnson, M.D.       | P                |
| 03     | Jeanine S. Huttner, M.D. | P                | 23     | Sanford Kimmel, M.D.      | P                |
| 04     | Francis Aona, M.D.       | P                | 24     | Harry Mayhew, M.D.        | S                |
| 05     | Janet Arrowsmith, M.D.   | P                | 25     | Leroy Rodgers, M.D.       | S                |
| 06     | David DeFrance, M.D.     | S                | 26     | Thomas Tafelski, M.D.     | S                |
| 07     | Judith Furlong, M.D.     | S                | 27     | Mark Zilkoski, M.D.       | G                |
| 08     | Leslie Jackson, M.D.     | G                | 28     | Ken Bertka, M.D.          | G                |
| 09     | Paul Langenkamp, M.D.    | S                | 29     | Mark DeMichiei, M.D.      | G                |
| 10     | Philip Lepkowski, M.D.   | S                | 30     | John Eggert, M.D.         | P                |
| 11     | Wendy Martin, M.D.       | S                | 31     | Jeanne Fiorito, M.D.      | P                |
| 12     | Denny Mauricio, M.D.     | P                | 32     | Michael Fitzpatrick, M.D. | P                |
| 13     | Hasmukh Parmar, M.D.     | P                | 33     | Charles Holt, D.O.        | P                |
| 14     | Ricardo Pena, M.D.       | P                | 34     | Richard Koby, M.D.        | P                |
| 15     | David Reames, M.D.       | P                | 35     | John Meier, M.D.          | P                |
| 16     | Ronald Reynolds, M.D.    | G                | 36     | Douglas Smucker, M.D.     | S                |
| 17     | Mark Steinmetz, M.D.     | G                | 37     | David Weldy, M.D.         | P                |
| 18     | Geza Torok, M.D.         | S                | 38     | Cheryl Zaborowski, M.D.   | P                |
| 19     | Mark Young, M.D.         | P                |        |                           |                  |

- a. The random numbers obtained from Appendix B.4 are 31, 94, 43, 36, 03, 24, 17, and 09. Which physicians should be contacted?
  - b. Select a random sample of four physicians using the random numbers from Appendix B.4.
  - c. Using systematic random sampling, every fifth physician is selected starting with the fourth physician in the list. Which physicians will be contacted?
  - d. Select a sample that includes two physicians in solo practice (S), two in partnership (P), and one in group practice (G). Explain your procedure.
- 21.** A population consists of the following three values: 1, 2, and 3.
- a. Sampling with replacement, list all possible samples of size 2 and compute the mean of every sample.
  - b. Find the means of the distribution of the sample mean and the population mean. Compare the two values.
  - c. Compare the dispersion of the population with that of the sample mean.
  - d. Describe the shapes of the two distributions.
- 22.** Based on all student records at Camford University, students spend an average of 5.5 hours per week playing organized sports. The population's standard deviation is 2.2 hours per week. Based on a sample of 121 students, Healthy Lifestyles Incorporated (HLI) would like to apply the central limit theorem to make various estimates.
- a. Compute the standard error of the sample mean.
  - b. What is the chance HLI will find a sample mean between 5 and 6 hours?
  - c. Calculate the probability that the sample mean will be between 5.3 and 5.7 hours.
  - d. How strange would it be to obtain a sample mean greater than 6.5 hours?
- 23.** eComputers Inc. recently completed the design for a new laptop model. Top management would like some assistance in pricing the new laptop. Two market research firms were contacted and asked to prepare a pricing strategy. Marketing-Gets-Results tested the new eComputers laptop with 50 randomly selected consumers who indicated they plan to purchase a laptop within the next year. The second marketing research firm, called Marketing-Reaps-Profits, test-marketed the new eComputers laptop with 200 current laptop owners. Which of the marketing research companies' test results will be more useful? Discuss why.
- 24.** Answer the following questions in one or two well-constructed sentences.
- a. What happens to the standard error of the mean if the sample size is increased?
  - b. What happens to the distribution of the sample means if the sample size is increased?
  - c. When using sample means to estimate the population mean, what is the benefit of using larger sample sizes?
- 25.** There are 25 motels in Goshen, Indiana. The number of rooms in each motel follows:

90 72 75 60 75 72 84 72 88 74 105 115 68 74 80 64 104 82 48 58 60 80 48 58 100

- a. Using a table of random numbers (Appendix B.4), select a random sample of five motels from this population.
  - b. Obtain a systematic sample by selecting a random starting point among the first five motels and then select every fifth motel.
  - c. Suppose the last five motels are "cut-rate" motels. Describe how you would select a random sample of three regular motels and two cut-rate motels.
- 26.** As a part of their customer service program, Global Airlines randomly selected 10 passengers from today's 9 a.m. Chicago–Tampa flight. Each sampled passenger will be interviewed about airport facilities, service, and so on. To select the sample, each passenger was given a number on boarding the aircraft. The numbers started with 001 and ended with 250.
- a. Select 10 usable numbers at random using Appendix B.4.
  - b. The sample of 10 could have been chosen using a systematic sample. Choose the first number using Appendix B.4, and then list the numbers to be interviewed.
  - c. Evaluate the two methods by giving the advantages and possible disadvantages.
  - d. What other way could a random sample be selected from the 250 passengers?

- 27.** Suppose your statistics instructor gave six examinations during the semester. You received the following exam scores (percent correct): 79, 64, 84, 82, 92, and 77. The instructor decided to randomly select two exam scores, compute their mean, and use this score to determine your final course grade.
- Compute the population mean.
  - How many different samples of two test grades are possible?
  - List all possible samples of size 2 and compute the mean of each.
  - Compute the mean of the sample means and compare it to the population mean.
  - If you were a student, would you like this arrangement? Would the result be different from dropping the lowest score? Write a brief report.
- 28.** At the downtown office of First National Bank, there are five tellers. Last week, the tellers made the following number of errors each: 2, 3, 5, 3, and 5.
- How many different samples of two tellers are possible?
  - List all possible samples of size 2 and compute the mean of each.
  - Compute the mean of the sample means and compare it to the population mean.
- 29.** The quality control department employs five technicians during the day shift. Listed below is the number of times each technician instructed the production foreman to shut down the manufacturing process last week.

| Technician | Shutdowns | Technician | Shutdowns |
|------------|-----------|------------|-----------|
| Taylor     | 4         | Rousche    | 3         |
| Hurley     | 3         | Huang      | 2         |
| Gupta      | 5         |            |           |

- How many different samples of two technicians are possible from this population?
  - List all possible samples of two observations each and compute the mean of each sample.
  - Compare the mean of the sample means with the population mean.
  - Compare the shape of the population distribution with the shape of the distribution of the sample means.
- 30.** The Appliance Center has six sales representatives at its North Jacksonville outlet. The following table lists the number of refrigerators sold by each representative last month.

| Sales Representative | Number Sold | Sales Representative | Number Sold |
|----------------------|-------------|----------------------|-------------|
| Zina Craft           | 54          | Jan Niles            | 48          |
| Woon Junge           | 50          | Molly Camp           | 50          |
| Ernie DeBrul         | 52          | Rachel Myak          | 52          |

- How many samples of size 2 are possible?
  - Select all possible samples of size 2 and compute the mean number sold.
  - Organize the sample means into a frequency distribution.
  - What is the mean of the population? What is the mean of the sample means?
  - What is the shape of the population distribution?
  - What is the shape of the distribution of the sample mean?
- 31.** Power + Inc. produces AA batteries used in remote-controlled toy cars. The mean life of these batteries follows the normal probability distribution with a mean of 35.0 hours and a standard deviation of 5.5 hours. As a part of its quality assurance program, Power + Inc. tests samples of 25 batteries.
- What can you say about the shape of the distribution of the sample mean?
  - What is the standard error of the distribution of the sample mean?
  - What proportion of the samples will have a mean useful life of more than 36 hours?
  - What proportion of the samples will have a mean useful life greater than 34.5 hours?

- e. What proportion of the samples will have a mean useful life between 34.5 and 36.0 hours?
- 32.** Majesty Video Production Inc. wants the mean length of its advertisements to be 30 seconds. Assume the distribution of ad length follows the normal distribution with a population standard deviation of 2 seconds. Suppose we select a sample of 16 ads produced by Majesty.
- What can we say about the shape of the distribution of the sample mean time?
  - What is the standard error of the mean time?
  - What percent of the sample means will be greater than 31.25 seconds?
  - What percent of the sample means will be greater than 28.25 seconds?
  - What percent of the sample means will be greater than 28.25 but less than 31.25 seconds?
- 33.** Recent studies indicate that the typical 50-year-old woman spends \$350 per year for personal-care products. The distribution of the amounts spent follows a normal distribution with a standard deviation of \$45 per year. We select a random sample of 40 women. The mean amount spent for those sampled is \$335. What is the likelihood of finding a sample mean this large or larger from the specified population?
- 34.** Information from the American Institute of Insurance indicates the mean amount of life insurance per household in the United States is \$165,000. This distribution follows the normal distribution with a standard deviation of \$40,000.
- If we select a random sample of 50 households, what is the standard error of the mean?
  - What is the expected shape of the distribution of the sample mean?
  - What is the likelihood of selecting a sample with a mean of at least \$167,000?
  - What is the likelihood of selecting a sample with a mean of more than \$155,000?
  - Find the likelihood of selecting a sample with a mean of more than \$155,000 but less than \$167,000.
- 35.** In the United States, the mean age of men when they marry for the first time follows the normal distribution with a mean of 29 years. The standard deviation of the distribution is 2.5 years. For a random sample of 60 men, what is the likelihood that the age when they were first married is less than 29.3 years?
- 36.** A recent study by the Greater Los Angeles Taxi Drivers Association showed that the mean fare charged for service from Hermosa Beach to Los Angeles International Airport is \$21 and the standard deviation is \$3.50. We select a sample of 15 fares.
- What is the likelihood that the sample mean is between \$20 and \$23?
  - What must you assume to make the above calculation?
- 37.** Crossett Trucking Company claims that the mean weight of its delivery trucks when they are fully loaded is 6,000 pounds and the standard deviation is 150 pounds. Assume that the population follows the normal distribution. Forty trucks are randomly selected and weighed. Within what limits will 95% of the sample means occur?
- 38.** The mean amount purchased by a typical customer at Churchill's Grocery Store is \$23.50, with a standard deviation of \$5.00. Assume the distribution of amounts purchased follows the normal distribution. For a sample of 50 customers, answer the following questions.
- What is the likelihood the sample mean is at least \$25.00?
  - What is the likelihood the sample mean is greater than \$22.50 but less than \$25.00?
  - Within what limits will 90% of the sample means occur?
- 39.** The mean performance score on a physical fitness test for Division I student-athletes is 947 with a standard deviation of 205. If you select a random sample of 60 of these students, what is the probability the mean is below 900?
- 40.** Suppose we roll a fair die two times.
- How many different samples are there?
  - List each of the possible samples and compute the mean.
  - On a chart similar to Chart 8–2, compare the distribution of sample means with the distribution of the population.
  - Compute the mean and the standard deviation of each distribution and compare them.



41. **FILE** Following is a list of the 50 states with the numbers 0 through 49 assigned to them.

| Number | State         | Number | State          |
|--------|---------------|--------|----------------|
| 0      | Alabama       | 25     | Montana        |
| 1      | Alaska        | 26     | Nebraska       |
| 2      | Arizona       | 27     | Nevada         |
| 3      | Arkansas      | 28     | New Hampshire  |
| 4      | California    | 29     | New Jersey     |
| 5      | Colorado      | 30     | New Mexico     |
| 6      | Connecticut   | 31     | New York       |
| 7      | Delaware      | 32     | North Carolina |
| 8      | Florida       | 33     | North Dakota   |
| 9      | Georgia       | 34     | Ohio           |
| 10     | Hawaii        | 35     | Oklahoma       |
| 11     | Idaho         | 36     | Oregon         |
| 12     | Illinois      | 37     | Pennsylvania   |
| 13     | Indiana       | 38     | Rhode Island   |
| 14     | Iowa          | 39     | South Carolina |
| 15     | Kansas        | 40     | South Dakota   |
| 16     | Kentucky      | 41     | Tennessee      |
| 17     | Louisiana     | 42     | Texas          |
| 18     | Maine         | 43     | Utah           |
| 19     | Maryland      | 44     | Vermont        |
| 20     | Massachusetts | 45     | Virginia       |
| 21     | Michigan      | 46     | Washington     |
| 22     | Minnesota     | 47     | West Virginia  |
| 23     | Mississippi   | 48     | Wisconsin      |
| 24     | Missouri      | 49     | Wyoming        |

- a. You wish to select a sample of eight from this list. The selected random numbers are 45, 15, 81, 09, 39, 43, 90, 26, 06, 45, 01, and 42. Which states are included in the sample?
- b. Select a systematic sample of every sixth item using the digit 02 as the starting point. Which states are included?
42. Human Resource Consulting (HRC) surveyed a random sample of 60 Twin Cities construction companies to find information on the costs of their health care plans. One of the items being tracked is the annual deductible that employees must pay. The Minnesota Department of Labor reports that historically the mean deductible amount per employee is \$502 with a standard deviation of \$100.
- a. Compute the standard error of the sample mean for HRC.
- b. What is the chance HRC finds a sample mean between \$477 and \$527?
- c. Calculate the likelihood that the sample mean is between \$492 and \$512.
- d. What is the probability the sample mean is greater than \$550?
43. Over the past decade, the mean number of hacking attacks experienced by members of the Information Systems Security Association is 510 per year with a standard deviation of 14.28 attacks. The number of attacks per year is normally distributed. Suppose nothing in this environment changes.
- a. What is the likelihood this group will suffer an average of more than 600 attacks in the next 10 years?
- b. Compute the probability the mean number of attacks over the next 10 years is between 500 and 600.
- c. What is the possibility they will experience an average of less than 500 attacks over the next 10 years?
44. An economist uses the price of a gallon of milk as a measure of inflation. She finds that the average price is \$3.82 per gallon and the population standard deviation is \$0.33. You decide to sample 40 convenience stores, collect their prices for a gallon of milk, and compute the mean price for the sample.

- a. What is the standard error of the mean in this experiment?
  - b. What is the probability that the sample mean is between \$3.78 and \$3.86?
  - c. What is the probability that the difference between the sample mean and the population mean is less than \$0.01?
  - d. What is the likelihood the sample mean is greater than \$3.92?
45. Nike's annual report says that the average American buys 6.5 pairs of sports shoes per year. Suppose a sample of 81 customers is surveyed and the population standard deviation of sports shoes purchased per year is 2.1.
- a. What is the standard error of the mean in this experiment?
  - b. What is the probability that the sample mean is between 6 and 7 pairs of sports shoes?
  - c. What is the probability that the difference between the sample mean and the population mean is less than 0.25 pair?
  - d. What is the likelihood the sample mean is greater than 7 pairs?

## DATA ANALYTICS

(The data for these exercises are available in Connect.)

46. **FILE** Refer to the North Valley Real Estate data, which report information on the homes sold last year. Assume the 105 homes is a population. Compute the population mean and the standard deviation of price. Select a sample of 10 homes. Compute the mean. Determine the likelihood of a sample mean price this high or higher.
47. **FILE** Refer to the Baseball 2016 data, which report information on the 30 Major League Baseball teams for the 2016 season. Over the last decade, the mean attendance per team followed a normal distribution with a mean of 2.45 million per team and a standard deviation of .71 million. Compute the mean attendance per team for the 2016 season. Determine the likelihood of a sample mean attendance this large or larger from the population.
48. **FILE** Refer to the Lincolnville School District bus data. Information provided by manufacturers of school buses suggests the mean maintenance cost per year is \$4,400 per bus with a standard deviation of \$1,000. Compute the mean maintenance cost for the Lincolnville buses. Do the Lincolnville data seem to be in line with those reported by the manufacturer? Specifically, what is the probability of Lincolnville's mean annual maintenance cost, or greater, given the manufacturer's data?

## PRACTICE TEST

### Part 1—Objective

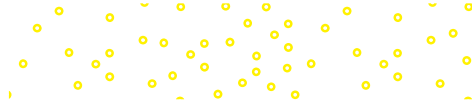
1. In a \_\_\_\_\_, each item in the population has the same chance of being included in the sample.
2. A sample should have at least how many observations? \_\_\_\_\_ (10, 30, 100, 1,000, no size restriction)
3. When a population is divided into groups based on some characteristic, such as region of the country, the groups are called \_\_\_\_\_.
4. The difference between a sample mean and the population mean is called the \_\_\_\_\_.
5. A probability distribution of all possible sample means for a particular sample size is the \_\_\_\_\_.
6. Suppose a population consisted of 10 individuals and we wished to list all possible samples of size 3. If sampling is without replacement, how many samples are there?
7. What is the name given to the standard deviation of the distribution of sample means? \_\_\_\_\_
8. The mean of all possible sample means is \_\_\_\_\_ the population mean. (always larger than, always smaller than, always equal to, not a constant relationship with)
9. If we increase the sample size from 10 to 20, the standard error of the mean will \_\_\_\_\_. (increase, decrease, stay the same, the result is not predictable)
10. If a population follows the normal distribution, what will be the shape of the distribution of sample means? \_\_\_\_\_

### Part 2—Problem

1. Americans spend a mean of 12.2 minutes per day in the shower. The distribution of time spent in the shower follows the normal distribution with a population standard deviation of 2.3 minutes. What is the likelihood that the mean time in the shower per day for a sample of 12 Americans is 11 minutes or less?

# 9

# Estimation and Confidence Intervals



©Jack Hollingsworth/Photodisc/Getty Images

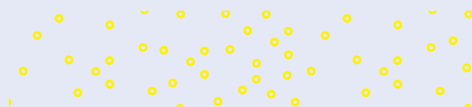
- ▲ **THE AMERICAN RESTAURANT ASSOCIATION** collected information on the number of meals eaten outside the home per week by young married couples. A survey of 60 couples showed the sample mean number of meals eaten outside the home was 2.76 meals per week, with a standard deviation of 0.75 meal per week. Construct a 99% confidence interval for the population mean. (See Exercise 32 and **LO9-2**.)

## LEARNING OBJECTIVES

---

*When you have completed this chapter, you will be able to:*

- LO9-1** Compute and interpret a point estimate of a population mean.
- LO9-2** Compute and interpret a confidence interval for a population mean.
- LO9-3** Compute and interpret a confidence interval for a population proportion.
- LO9-4** Calculate the required sample size to estimate a population proportion or population mean.



**STATISTICS IN ACTION**

On all new cars, a fuel economy estimate is prominently displayed on the window sticker as required by the Environmental Protection Agency (EPA). Often, fuel economy is a factor in a consumer's choice of a new car because of fuel costs or environmental concerns. The fuel estimates for a 2017 BMW 328i Sedan (4-cylinder, automatic) are 32 miles per gallon (mpg) in the city and 42 on the highway. The EPA recognizes that actual fuel economy may differ from the estimates by noting, "No test can simulate all possible combinations of conditions and climate, driver behavior, and car care habits. Actual mileage depends on how, when, and where the vehicle is driven. The EPA has found that the mpg obtained by most drivers will be within a few mpg of the estimates."

**LO9-1**

Compute and interpret a point estimate of a population mean.

**INTRODUCTION**

The previous chapter began our discussion of sampling. We introduced both the reasons for, and the methods of, sampling. The reasons for sampling were:

- Contacting the entire population is too time-consuming.
- Studying all the items in the population is often too expensive.
- The sample results are usually adequate.
- Certain tests are destructive.
- Checking all the items is physically impossible.

There are several methods of sampling. Simple random sampling is the most widely used method. With this type of sampling, each member of the population has the same chance of being selected to be a part of the sample. Other methods of sampling include systematic sampling, stratified sampling, and cluster sampling.

Chapter 8 assumes information about the population, such as the mean, the standard deviation, or the shape of the population, is known. In most business situations, such information is not available. In fact, one purpose of sampling is to estimate some of these values. For example, you select a sample from a population and use the mean of the sample to estimate the mean of the population.

This chapter considers several important aspects of sampling. We begin by studying **point estimates**. A point estimate is a single value (point) computed from sample information and used to estimate a population value. For example, we may be interested in the number of hours worked by consultants employed by Boston Consulting Group. Using simple random sampling, we select 50 consultants and ask each of them how many hours they worked last week. The sample's mean is a point estimate of the unknown population mean. A more informative approach is to present a range of values where we expect the population parameter to occur. Such a range of values is called a **confidence interval**.

When sampling from a population, an important decision is to determine the size of a sample. How many voters should a polling organization contact to forecast the election outcome? How many products do we need to examine to ensure our quality level? This chapter also develops a strategy for determining the appropriate number of observations in the sample.

**POINT ESTIMATE FOR A POPULATION MEAN**

A point estimate is a single statistic used to estimate a population parameter. Suppose Best Buy Inc. wants to estimate the mean age of people who purchase LCD HDTV televisions. They select a random sample of 75 recent purchases, determine the age of each buyer, and compute the mean age of the buyers in the sample. The mean of this sample is a **point estimate** of the population mean.

**POINT ESTIMATE** The statistic, computed from sample information, that estimates a population parameter.

The following examples illustrate point estimates of population means.

1. Tourism is a major source of income for many Caribbean countries, such as Barbados. Suppose the Bureau of Tourism for Barbados wants an estimate of the mean amount spent by tourists visiting the country. It would not be feasible to contact each tourist. Therefore, 500 tourists are randomly selected as they depart the country and asked in detail about their spending while visiting Barbados. The mean amount spent by the sample of 500 tourists is an estimate of the unknown population parameter. That is, we let the sample mean serve as a point estimate of the population mean.

2. Litchfield Home Builders Inc. builds homes in the southeastern region of the United States. One of the major concerns of new buyers is the date when the home will be completed. Recently, Litchfield has been telling customers, “Your home will be completed 45 working days from the date we begin installing dry-wall.” The customer relations department at Litchfield wishes to compare this pledge with recent experience. A sample of 50 homes completed this year revealed that the point estimate of the population mean is 46.7 working days from the start of drywall to the completion of the home. Is it reasonable to conclude that the population mean is still 45 days and that the difference between the sample mean (46.7 days) and the proposed population mean (45 days) is sampling error? In other words, is the sample mean significantly different from the population mean?



©Andersen Ross/Getty Images RF

3. Recent medical studies indicate that exercise is an important part of a person’s overall health. The director of human resources at OCF, a large glass manufacturer, wants an estimate of the number of hours per week employees spend exercising. A sample of 70 employees reveals the mean number of hours of exercise last week is 3.3. This value is a point estimate of the unknown population mean.

The sample mean,  $\bar{x}$ , is not the only point estimate of a population parameter. For example,  $p$ , a sample proportion, is a point estimate of  $\pi$ , the population proportion; and  $s$ , the sample standard deviation, is a point estimate of  $\sigma$ , the population standard deviation.

### LO9-2

Compute and interpret a confidence interval for a population mean.

## CONFIDENCE INTERVALS FOR A POPULATION MEAN

A point estimate, however, tells only part of the story. While we expect the point estimate to be close to the population parameter, we would like to measure how close it really is. A confidence interval serves this purpose. For example, we estimate the mean yearly income for construction workers in the New York–New Jersey area is \$85,000. The range of this estimate might be from \$81,000 to \$89,000. We can describe how confident we are that the population parameter is in the interval. We might say, for instance, that we are 90% confident that the mean yearly income of construction workers in the New York–New Jersey area is between \$81,000 and \$89,000.

**CONFIDENCE INTERVAL** A range of values constructed from sample data so that the population parameter is likely to occur within that range at a specified probability. The specified probability is called the *level of confidence*.

To compute a confidence interval for a population mean, we will consider two situations:

- We use sample data to estimate  $\mu$  with  $\bar{x}$  and the population standard deviation ( $\sigma$ ) is known.
- We use sample data to estimate  $\mu$  with  $\bar{x}$  and the population standard deviation is unknown. In this case, we substitute the sample standard deviation ( $s$ ) for the population standard deviation ( $\sigma$ ).

There are important distinctions in the assumptions between these two situations. We first consider the case where  $\sigma$  is known.

### Population Standard Deviation, Known $\sigma$

A confidence interval is computed using two statistics: the sample mean,  $\bar{x}$ , and the standard deviation. From previous chapters, you know that the standard deviation is an important statistic because it measures the dispersion, or variation, of a population or

sampling distribution. In computing a confidence interval, the standard deviation is used to compute the limits of the confidence interval.

To demonstrate the idea of a confidence interval, we start with one simplifying assumption. That assumption is that we know the value of the population standard deviation,  $\sigma$ . Typically, we know the population standard deviation in situations where we have a long history of collected data. Examples are data from monitoring processes that fill soda bottles or cereal boxes, and the results of the SAT Reasoning Test (for college admission). Knowing  $\sigma$  allows us to simplify the development of a confidence interval because we can use the standard normal distribution from Chapter 7.

Recall that the sampling distribution of the sample mean is the distribution of all sample means,  $\bar{x}$ , of sample size  $n$  from a population. The population standard deviation,  $\sigma$ , is known. From this information, and the central limit theorem, we know that the sampling distribution follows the normal probability distribution with a mean of  $\mu$  and a standard deviation  $\sigma/\sqrt{n}$ . Also recall that this value is called the standard error.

The results of the central limit theorem allow us to make the following general confidence interval statements using z-statistics:

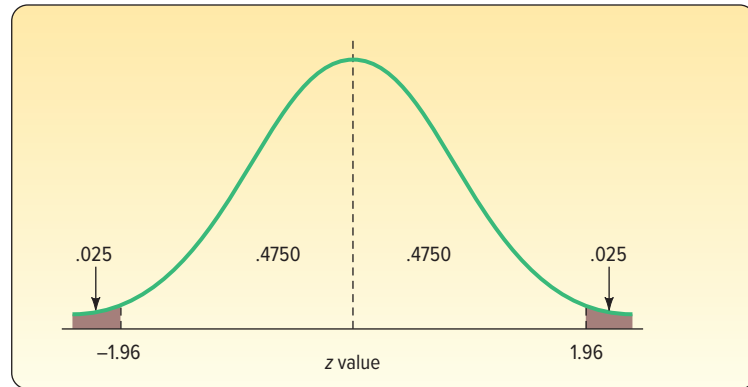
1. Ninety-five percent of all confidence intervals computed from random samples selected from a population will contain the population mean. These intervals are computed using a z-statistic equal to 1.96.
2. Ninety percent of all confidence intervals computed from random samples selected from a population will contain the population mean. These confidence intervals are computed using a z-statistic equal to 1.65.

These confidence interval statements provide examples of *levels of confidence* and are called a **95% confidence interval** and a **90% confidence interval**. The 95% and 90% are the levels of confidence and refer to the percentage of similarly constructed intervals that would include the parameter being estimated—in this case,  $\mu$ , the population mean.

How are the values of 1.96 and 1.65 obtained? First, let's look for the z value for a 95% confidence interval. The following diagram and Table 9–1 will help explain. Table 9–1 is a reproduction of the standard normal table in Appendix B.3 However, many rows and columns have been eliminated to allow us to better focus on particular rows and columns.

1. First, we divide the confidence level in half, so  $.9500/2 = .4750$ .
2. Next, we find the value .4750 in the body of Table 9–1. Note that .4750 is located in the table at the intersection of a row and a column.
3. Locate the corresponding row value in the left margin, which is 1.9, and the column value in the top margin, which is .06. Adding the row and column values gives us a z value of 1.96.
4. Thus, the probability of finding a z value between 0 and 1.96 is .4750.
5. Likewise, because the normal distribution is symmetric, the probability of finding a z value between  $-1.96$  and 0 is also .4750.
6. When we add these two probabilities, the probability that a z value is between  $-1.96$  and 1.96 is .9500.

For the 90% level of confidence, we follow the same steps. First, one-half of the desired confidence interval is .4500. A search of Table 9–1 does not reveal this exact value. However, it is between two values, .4495 and .4505. As in step three, we locate each value in the table. The first, .4495, corresponds to a z value of 1.64 and the second, .4505, corresponds to a z value of 1.65. To be conservative, we will select the larger of the two z values, 1.65, and the exact level of confidence is 90.1%, or  $2(0.4505)$ . Next, the probability of finding a z value between  $-1.65$  and 0 is .4505, and the probability that a z value is between  $-1.65$  and 1.65 is .9010.



**TABLE 9–1** The Standard Normal Table for Selected Values

| z   | 0.00   | 0.01   | 0.02   | 0.03   | 0.04   | 0.05   | 0.06   | 0.07   |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|
| ⋮   | ⋮      | ⋮      | ⋮      | ⋮      | ⋮      | ⋮      | ⋮      | ⋮      |
| 1.5 | 0.4332 | 0.4345 | 0.4357 | 0.4370 | 0.4382 | 0.4394 | 0.4406 | 0.4418 |
| 1.6 | 0.4452 | 0.4463 | 0.4474 | 0.4484 | 0.4495 | 0.4505 | 0.4515 | 0.4525 |
| 1.7 | 0.4554 | 0.4564 | 0.4573 | 0.4582 | 0.4591 | 0.4599 | 0.4608 | 0.4616 |
| 1.8 | 0.4641 | 0.4649 | 0.4656 | 0.4664 | 0.4671 | 0.4678 | 0.4686 | 0.4693 |
| 1.9 | 0.4713 | 0.4719 | 0.4726 | 0.4732 | 0.4738 | 0.4744 | 0.4750 | 0.4756 |
| 2.0 | 0.4772 | 0.4778 | 0.4783 | 0.4788 | 0.4793 | 0.4798 | 0.4803 | 0.4808 |
| 2.1 | 0.4821 | 0.4826 | 0.4830 | 0.4834 | 0.4838 | 0.4842 | 0.4846 | 0.4850 |
| 2.2 | 0.4861 | 0.4864 | 0.4868 | 0.4871 | 0.4875 | 0.4878 | 0.4881 | 0.4884 |

How do we determine a 95% confidence interval? The width of the interval is determined by two factors: (1) the level of confidence, as described in the previous section, and (2) the size of the standard error of the mean. To find the standard error of the mean, recall from the previous chapter [see formula (8–1) on page 231] that the standard error of the mean reports the variation in the distribution of sample means. It is really the standard deviation of the distribution of sample means. The formula is repeated below:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where:

$\sigma_{\bar{x}}$  is the symbol for the standard error of the mean. We use a Greek letter because it is a population value, and the subscript  $\bar{x}$  reminds us that it refers to a sampling distribution of the sample means.

$\sigma$  is the population standard deviation.

$n$  is the number of observations in the sample.

The size of the standard error is affected by two values. The first is the standard deviation of the population. The larger the population standard deviation,  $\sigma$ , the larger  $\sigma/\sqrt{n}$ . If the population is homogeneous, resulting in a small population standard deviation, the standard error will also be small. However, the standard error is also affected by the number of observations in the sample. A large number of observations in the sample will result in a small standard error of estimate, indicating that there is less variability in the sample means.

We can summarize the calculation for a 95% confidence interval using the following formula:

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

Similarly, a 90.1% confidence interval is computed as follows:

$$\bar{x} \pm 1.65 \frac{\sigma}{\sqrt{n}}$$

The values 1.96 and 1.65 are z values corresponding to the 95% and the 90.1% confidence intervals, respectively. However, we are not restricted to these values. We can select any confidence level between 0 and 100% and find the corresponding value for z. In general, a confidence interval for the population mean when the population follows the normal distribution and the population standard deviation is known is computed by:

**CONFIDENCE INTERVAL FOR A  
POPULATION MEAN WITH  $\sigma$  KNOWN**

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}} \quad (9-1)$$



©Chad Zuber/123RF

To explain these ideas, consider the following example. Del Monte Foods distributes diced peaches in 4.5-ounce plastic cups. To ensure that each cup contains at least the required amount, Del Monte sets the filling operation to dispense 4.51 ounces of peaches and gel in each cup. Of course, not every cup will contain exactly 4.51 ounces of peaches and gel. Some cups will have more and others less. From historical data, Del Monte knows that 0.04 ounce is the standard deviation of the filling process and that the amount, in ounces, follows the normal probability distribution. The quality control technician selects a sample of 64 cups at the start of each shift, measures the amount in each cup, computes the mean fill amount, and then develops a 95% confidence interval for the population mean. Using the confidence interval, is the process filling the cups to the desired amount? This morning's sample of 64 cups had a sample mean of 4.507 ounces. Based on this information, the 95% confidence interval is:

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 4.507 \pm 1.96 \frac{0.04}{\sqrt{64}} = 4.507 \pm 0.0098$$

The 95% confidence interval estimates that the population mean is between 4.4972 ounces and 4.5168 ounces of peaches and gel. Recall that the process is set to fill each cup with 4.51 ounces. Because the desired fill amount of 4.51 ounces is in this interval, we conclude that the filling process is achieving the desired results. In other words, it is reasonable to conclude that the sample mean of 4.507 could have come from a population distribution with a mean of 4.51 ounces.

In this example, we observe that the population mean of 4.51 ounces is in the confidence interval. But this is not always the case. If we selected 100 samples of 64 cups from the population, calculated the sample mean, and developed a confidence interval based on each sample, we would expect to find the population mean in about 95 of the 100 intervals. Or, in contrast, about five of the intervals would not contain the population mean. From Chapter 8, this is called sampling error. The following example details repeated sampling from a population.

**EXAMPLE**

The American Management Association (AMA) is studying the income of store managers in the retail industry. A random sample of 49 managers reveals a sample mean of \$45,420. The standard deviation of this population is \$2,050. The association would like answers to the following questions:

1. What is the population mean?
2. What is a reasonable range of values for the population mean?
3. How do we interpret these results?



### SOLUTION

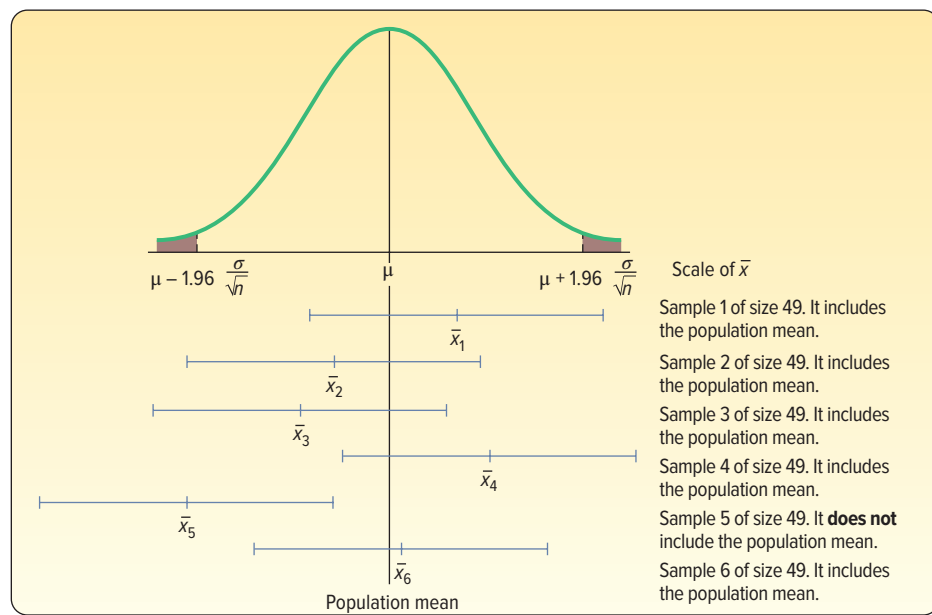
Generally, distributions of salary and income are positively skewed because a few individuals earn considerably more than others, thus skewing the distribution in the positive direction. Fortunately, the central limit theorem states that the sampling distribution of the mean becomes a normal distribution as sample size increases. In this instance, a sample of 49 store managers is large enough that we can assume that the sampling distribution will follow the normal distribution. Now to answer the questions posed in the example.

1. **What is the population mean?** In this case, we do not know. We do know the sample mean is \$45,420. Hence, our best estimate of the unknown population value is the corresponding sample statistic. Thus, the sample mean of \$45,420 is a *point estimate* of the unknown population mean.
2. **What is a reasonable range of values for the population mean?** The AMA decides to use the 95% level of confidence. To determine the corresponding confidence interval, we use formula (9–1):

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}} = \$45,420 \pm 1.96 \frac{\$2,050}{\sqrt{49}} = \$45,420 \pm \$574$$

The confidence interval limits are \$44,846 and \$45,994 determined by subtracting \$574 and adding \$574 to the sample mean. The degree or level of confidence is 95% and the confidence interval is from \$44,846 to \$45,994. The value, \$574, is called the margin of error.

3. **How do we interpret these results?** Suppose we select many samples of 49 store managers, perhaps several hundred. For each sample, we compute the mean and then construct a 95% confidence interval, such as we did in the previous section. We could expect about 95% of these confidence intervals to contain the *population* mean. About 5% of the intervals would not contain the population mean annual income, which is  $\mu$ . However, a particular confidence interval either contains the population parameter or it does not. The following diagram shows the results of selecting samples from the population of store managers in the retail industry, computing the mean of each, and then, using formula (9–1), determining a 95% confidence interval for the population mean. Note that not all intervals include the population mean. Both the endpoints of the fifth sample are less than the population mean. We attribute this to sampling error, and it is the risk we assume when we select the level of confidence.



## A Computer Simulation

With statistical software, we can create random samples of a desired sample size,  $n$ , from a population. For each sample of  $n$  observations with corresponding numerical values, we can calculate the sample mean. With the sample mean, population standard deviation, and confidence level, we can determine the confidence interval for each sample. Then, using all samples and the confidence intervals, we can find the frequency that the population mean is included in the confidence intervals. The following example does just that.

### EXAMPLE

From many years in the automobile leasing business, Town Bank knows that the mean distance driven on an automobile with a four-year lease is 50,000 miles and the standard deviation is 5,000 miles. These are population values. Suppose Town Bank would like to experiment with the idea of sampling to estimate the population mean of 50,000 miles. Town Bank decides to choose a sample size of 30 observations and a 95% confidence interval to estimate the population mean. Based on the experiment, we want to count the number of confidence intervals that include the population mean of 50,000. We expect about 95%, or 57 of the 60 intervals, will include the population mean. To make the calculations easier to understand, we'll conduct the study in thousands of miles, instead of miles.

### SOLUTION

Using statistical software, 60 random samples of 30 observations,  $n = 30$ , are generated and the sample means for each sample computed. Then, using the  $n$  of 30 and a standard error of 0.913 ( $\sigma/\sqrt{n} = 5/\sqrt{30}$ ), a 95% confidence interval is computed for each sample. The results of the experiment are shown next.

| Sample | Sample Observations |    |    |    |    |   |   |   |    |    | Sample Mean | 95% Confidence Limits |    |       |             |             |
|--------|---------------------|----|----|----|----|---|---|---|----|----|-------------|-----------------------|----|-------|-------------|-------------|
|        | 1                   | 2  | 3  | 4  | 5  | – | – | – | 26 | 27 |             | 28                    | 29 | 30    | Lower Limit | Upper Limit |
| 1      | 56                  | 47 | 47 | 48 | 58 | – | – | – | 55 | 62 | 48          | 61                    | 57 | 51.6  | 49.811      | 53.389      |
| 2      | 55                  | 51 | 52 | 40 | 53 | – | – | – | 47 | 54 | 55          | 55                    | 45 | 50.77 | 48.981      | 52.559      |
| 3      | 42                  | 46 | 48 | 46 | 41 | – | – | – | 50 | 52 | 50          | 47                    | 45 | 48.63 | 46.841      | 50.419      |
| 4      | 52                  | 49 | 55 | 47 | 49 | – | – | – | 46 | 56 | 49          | 43                    | 50 | 49.9  | 48.111      | 51.689      |
| 5      | 48                  | 50 | 53 | 48 | 45 | – | – | – | 46 | 51 | 61          | 49                    | 47 | 49.03 | 47.241      | 50.819      |
| 6      | 49                  | 44 | 47 | 46 | 48 | – | – | – | 51 | 44 | 51          | 52                    | 43 | 47.73 | 45.941      | 49.519      |
| 7      | 50                  | 53 | 39 | 50 | 46 | – | – | – | 55 | 47 | 43          | 50                    | 57 | 50.2  | 48.411      | 51.989      |
| 8      | 47                  | 51 | 49 | 58 | 44 | – | – | – | 49 | 57 | 54          | 48                    | 48 | 51.17 | 49.381      | 52.959      |
| 9      | 51                  | 44 | 47 | 56 | 45 | – | – | – | 45 | 51 | 49          | 49                    | 52 | 50.33 | 48.541      | 52.119      |
| 10     | 45                  | 44 | 52 | 52 | 56 | – | – | – | 52 | 51 | 52          | 50                    | 48 | 50    | 48.211      | 51.789      |
| 11     | 43                  | 52 | 54 | 46 | 54 | – | – | – | 43 | 46 | 49          | 52                    | 52 | 51.2  | 49.411      | 52.989      |
| 12     | 57                  | 53 | 48 | 42 | 55 | – | – | – | 49 | 44 | 46          | 46                    | 48 | 49.8  | 48.011      | 51.589      |
| 13     | 53                  | 39 | 47 | 51 | 53 | – | – | – | 42 | 44 | 44          | 55                    | 58 | 49.6  | 47.811      | 51.389      |
| 14     | 56                  | 55 | 45 | 43 | 57 | – | – | – | 48 | 51 | 52          | 55                    | 47 | 49.03 | 47.241      | 50.819      |
| 15     | 49                  | 50 | 39 | 45 | 44 | – | – | – | 49 | 43 | 44          | 51                    | 51 | 49.37 | 47.581      | 51.159      |
| 16     | 46                  | 44 | 55 | 53 | 55 | – | – | – | 44 | 53 | 53          | 43                    | 44 | 50.13 | 48.341      | 51.919      |
| 17     | 64                  | 52 | 55 | 55 | 43 | – | – | – | 58 | 46 | 52          | 58                    | 55 | 52.47 | 50.681      | 54.259      |
| 18     | 57                  | 51 | 60 | 40 | 53 | – | – | – | 50 | 51 | 53          | 46                    | 52 | 50.1  | 48.311      | 51.889      |
| 19     | 50                  | 49 | 51 | 57 | 45 | – | – | – | 53 | 52 | 40          | 45                    | 52 | 49.6  | 47.811      | 51.389      |
| 20     | 45                  | 46 | 53 | 57 | 49 | – | – | – | 49 | 43 | 43          | 53                    | 48 | 49.47 | 47.681      | 51.259      |
| 21     | 52                  | 45 | 51 | 52 | 45 | – | – | – | 43 | 49 | 49          | 58                    | 53 | 50.43 | 48.641      | 52.219      |
| 22     | 48                  | 48 | 52 | 49 | 40 | – | – | – | 50 | 47 | 54          | 51                    | 45 | 47.53 | 45.741      | 49.319      |

(continued)

| Sample | Sample Observations |    |    |    |    |   |   |   |    |    |    |    |    |             | Sample Mean | 95% Confidence Limits |  |
|--------|---------------------|----|----|----|----|---|---|---|----|----|----|----|----|-------------|-------------|-----------------------|--|
|        | 1                   | 2  | 3  | 4  | 5  | - | - | - | 26 | 27 | 28 | 29 | 30 | Lower Limit |             | Upper Limit           |  |
| 23     | 48                  | 50 | 50 | 53 | 44 | - | - | - | 48 | 57 | 52 | 44 | 39 | 49.1        | 47.311      | 50.889                |  |
| 24     | 51                  | 51 | 40 | 54 | 52 | - | - | - | 54 | 45 | 50 | 57 | 48 | 50.13       | 48.341      | 51.919                |  |
| 25     | 48                  | 63 | 41 | 52 | 41 | - | - | - | 48 | 50 | 48 | 44 | 53 | 49.33       | 47.541      | 51.119                |  |
| 26     | 47                  | 45 | 48 | 59 | 49 | - | - | - | 44 | 47 | 49 | 55 | 42 | 49.63       | 47.841      | 51.419                |  |
| 27     | 52                  | 45 | 60 | 51 | 52 | - | - | - | 52 | 50 | 54 | 46 | 52 | 49.4        | 47.611      | 51.189                |  |
| 28     | 46                  | 48 | 46 | 57 | 51 | - | - | - | 51 | 50 | 51 | 41 | 52 | 49.33       | 47.541      | 51.119                |  |
| 29     | 46                  | 48 | 45 | 42 | 48 | - | - | - | 49 | 43 | 59 | 46 | 50 | 48.27       | 46.481      | 50.059                |  |
| 30     | 55                  | 48 | 47 | 48 | 48 | - | - | - | 47 | 59 | 54 | 51 | 42 | 50.53       | 48.741      | 52.319                |  |
| 31     | 58                  | 49 | 56 | 46 | 46 | - | - | - | 44 | 51 | 47 | 51 | 46 | 50.77       | 48.981      | 52.559                |  |
| 32     | 53                  | 54 | 52 | 58 | 55 | - | - | - | 53 | 52 | 45 | 44 | 51 | 50          | 48.211      | 51.789                |  |
| 33     | 50                  | 57 | 56 | 51 | 51 | - | - | - | 58 | 47 | 50 | 56 | 46 | 49.7        | 47.911      | 51.489                |  |
| 34     | 61                  | 48 | 49 | 53 | 54 | - | - | - | 46 | 46 | 56 | 45 | 54 | 50.03       | 48.241      | 51.819                |  |
| 35     | 43                  | 42 | 43 | 46 | 49 | - | - | - | 49 | 49 | 56 | 51 | 45 | 49.43       | 47.641      | 51.219                |  |
| 36     | 39                  | 48 | 48 | 51 | 44 | - | - | - | 54 | 52 | 47 | 50 | 52 | 50.07       | 48.281      | 51.859                |  |
| 37     | 48                  | 43 | 57 | 42 | 54 | - | - | - | 52 | 50 | 59 | 50 | 52 | 50.17       | 48.381      | 51.959                |  |
| 38     | 55                  | 43 | 49 | 57 | 45 | - | - | - | 41 | 51 | 51 | 52 | 52 | 49.5        | 47.711      | 51.289                |  |
| 39     | 47                  | 49 | 58 | 54 | 54 | - | - | - | 50 | 56 | 51 | 56 | 58 | 50.37       | 48.581      | 52.159                |  |
| 40     | 47                  | 56 | 41 | 50 | 54 | - | - | - | 46 | 56 | 61 | 61 | 45 | 51.6        | 49.811      | 53.389                |  |
| 41     | 48                  | 47 | 42 | 47 | 62 | - | - | - | 44 | 47 | 49 | 55 | 43 | 49.43       | 47.641      | 51.219                |  |
| 42     | 46                  | 49 | 43 | 36 | 52 | - | - | - | 45 | 51 | 46 | 51 | 43 | 47.67       | 45.881      | 49.459                |  |
| 43     | 44                  | 48 | 49 | 48 | 51 | - | - | - | 47 | 52 | 51 | 48 | 49 | 49.63       | 47.841      | 51.419                |  |
| 44     | 45                  | 52 | 54 | 54 | 49 | - | - | - | 49 | 45 | 53 | 50 | 52 | 49.07       | 47.281      | 50.859                |  |
| 45     | 54                  | 46 | 54 | 45 | 48 | - | - | - | 55 | 38 | 56 | 50 | 62 | 49.53       | 47.741      | 51.319                |  |
| 46     | 48                  | 50 | 49 | 52 | 51 | - | - | - | 53 | 57 | 58 | 46 | 50 | 49.9        | 48.111      | 51.689                |  |
| 47     | 54                  | 55 | 46 | 55 | 50 | - | - | - | 56 | 54 | 50 | 55 | 51 | 50.5        | 48.711      | 52.289                |  |
| 48     | 45                  | 47 | 47 | 63 | 44 | - | - | - | 45 | 53 | 42 | 53 | 50 | 50.1        | 48.311      | 51.889                |  |
| 49     | 47                  | 47 | 48 | 54 | 56 | - | - | - | 50 | 48 | 54 | 49 | 51 | 49.93       | 48.141      | 51.719                |  |
| 50     | 45                  | 61 | 51 | 45 | 54 | - | - | - | 55 | 52 | 47 | 45 | 53 | 51.03       | 49.241      | 52.819                |  |
| 51     | 49                  | 62 | 43 | 49 | 48 | - | - | - | 49 | 58 | 42 | 58 | 52 | 51.07       | 49.281      | 52.859                |  |
| 52     | 54                  | 52 | 62 | 43 | 54 | - | - | - | 51 | 57 | 49 | 58 | 55 | 50.17       | 48.381      | 51.959                |  |
| 53     | 46                  | 50 | 59 | 56 | 46 | - | - | - | 50 | 51 | 52 | 54 | 53 | 50.47       | 48.681      | 52.259                |  |
| 54     | 52                  | 50 | 48 | 48 | 58 | - | - | - | 58 | 52 | 43 | 61 | 54 | 51.77       | 49.981      | 53.559                |  |
| 55     | 45                  | 44 | 46 | 56 | 46 | - | - | - | 43 | 45 | 63 | 48 | 56 | 49.37       | 47.581      | 51.159                |  |
| 56     | 60                  | 50 | 56 | 51 | 43 | - | - | - | 45 | 43 | 49 | 59 | 54 | 50.37       | 48.581      | 52.159                |  |
| 57     | 59                  | 56 | 43 | 47 | 52 | - | - | - | 49 | 54 | 50 | 50 | 57 | 49.53       | 47.741      | 51.319                |  |
| 58     | 52                  | 55 | 48 | 51 | 40 | - | - | - | 53 | 51 | 51 | 52 | 47 | 49.77       | 47.981      | 51.559                |  |
| 59     | 53                  | 50 | 44 | 53 | 52 | - | - | - | 47 | 50 | 55 | 46 | 51 | 50.07       | 48.281      | 51.859                |  |
| 60     | 55                  | 54 | 50 | 52 | 43 | - | - | - | 57 | 50 | 48 | 47 | 53 | 52.07       | 50.281      | 53.859                |  |

To explain, in the first row, the statistical software computed 30 random observations from a population distribution with a mean of 50 and a standard deviation of 5. To conserve space, only observations 1 through 5 and 26 through 30 are listed. The first sample's mean is computed and listed as 51.6. In the next columns, the upper and lower limits of the 95% confidence interval for the first sample are shown. The confidence interval calculation for the first sample follows:

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 51.6 \pm 1.96 \frac{5}{\sqrt{30}} = 51.6 \pm 1.789$$

This calculation is repeated for all samples. The results of the experiment show that 93.33%, or 56 of the 60 confidence intervals, include the population mean of 50. 93.33% is close to the estimate that 95%, or 57, of the intervals will include the population mean. Using the complement, we expected 5%, or three, of the intervals would not include the population mean. The experiment resulted in 6.67%, or four, of

the 60 intervals that did not include the population mean. The particular intervals, 6, 17, 22, and 42, are highlighted in yellow. This is another example of sampling error, or the possibility that a particular random sample may not be a good representation of the population. In each of these four samples, the mean of the sample is either much less or much more than the population mean. Because of random sampling, the mean of the sample is not a good estimate of the population mean, and the confidence interval based on the sample's mean does not include the population mean.

## SELF-REVIEW 9-1



The Bun-and-Run is a franchise fast-food restaurant located in the Northeast specializing in half-pound hamburgers, fish sandwiches, and chicken sandwiches. Soft drinks and French fries also are available. The marketing department of Bun-and-Run Inc. reports that the distribution of daily sales for their restaurants follows the normal distribution and that the population standard deviation is \$3,000. A sample of 40 franchises showed the mean daily sales to be \$20,000.

- What is the population mean of daily sales for Bun-and-Run franchises?
- What is the best estimate of the population mean? What is this value called?
- Develop a 95% confidence interval for the population mean of daily sales.
- Interpret the confidence interval.

## EXERCISES

- A sample of 49 observations is taken from a normal population with a standard deviation of 10. The sample mean is 55. Determine the 99% confidence interval for the population mean.
- A sample of 81 observations is taken from a normal population with a standard deviation of 5. The sample mean is 40. Determine the 95% confidence interval for the population mean.
- A sample of 250 observations is selected from a normal population with a population standard deviation of 25. The sample mean is 20.
  - Determine the standard error of the mean.
  - Explain why we can use formula (9-1) to determine the 95% confidence interval.
  - Determine the 95% confidence interval for the population mean.
- Suppose you know  $\sigma$  and you want an 85% confidence level. What value would you use as  $z$  in formula (9-1)?
- A research firm conducted a survey of 49 randomly selected Americans to determine the mean amount spent on coffee during a week. The sample mean was \$20 per week. The population distribution is normal with a standard deviation of \$5.
  - What is the point estimate of the population mean? Explain what it indicates.
  - Using the 95% level of confidence, determine the confidence interval for  $\mu$ . Explain what it indicates.
- Refer to the previous exercise. Instead of 49, suppose that 64 Americans were surveyed about their weekly expenditures on coffee. Assume the sample mean remained the same.
  - What is the 95% confidence interval estimate of  $\mu$ ?
  - Explain why this confidence interval is narrower than the one determined in the previous exercise.
- Bob Nale is the owner of Nale's Quick Fill. Bob would like to estimate the mean number of gallons of gasoline sold to his customers. Assume the number of gallons sold follows the normal distribution with a population standard deviation of 2.30 gallons. From his records, he selects a random sample of 60 sales and finds the mean number of gallons sold is 8.60.
  - What is the point estimate of the population mean?
  - Develop a 99% confidence interval for the population mean.
  - Interpret the meaning of part (b).

8. Dr. Patton is a professor of English. Recently she counted the number of misspelled words in a group of student essays. She noted the distribution of misspelled words per essay followed the normal distribution with a population standard deviation of 2.44 words per essay. For her 10 a.m. section of 40 students, the mean number of misspelled words was 6.05. Construct a 95% confidence interval for the mean number of misspelled words in the population of student essays.

## Population Standard Deviation, $\sigma$ Unknown

In the previous section, we assumed the population standard deviation was known. In the case involving Del Monte 4.5-ounce cups of peaches, there would likely be a long history of measurements in the filling process. Therefore, it is reasonable to assume the standard deviation of the population is available. However, in most sampling situations the population standard deviation ( $\sigma$ ) is not known. Here are some examples where we wish to estimate the population means and it is unlikely we would know the population standard deviations. Suppose each of these studies involves students at West Virginia University.

- The Dean of the Business College wants to estimate the mean number of hours full-time students work at paying jobs each week. He selects a sample of 30 students, contacts each student, and asks them how many hours they worked last week. From the sample information, he can calculate the sample mean, but it is not likely he would know or be able to find the *population* standard deviation ( $\sigma$ ) required in formula (9–1).
- The Dean of Students wants to estimate the distance the typical commuter student travels to class. She selects a sample of 40 commuter students, contacts each, and determines the one-way distance from each student's home to the center of campus. From the sample data, she calculates the mean travel distance, that is,  $\bar{x}$ . It is unlikely the standard deviation of the population would be known or available, again making formula (9–1) unusable.
- The Director of Student Loans wants to estimate the mean amount owed on student loans at the time of graduation. The director selects a sample of 20 graduating students and contacts each to find the information. From the sample information, the director can estimate the mean amount. However, to develop a confidence interval using formula (9–1), the population standard deviation is necessary. It is not likely this information is available.

Fortunately we can use the sample standard deviation to estimate the population standard deviation. That is, we use  $s$ , the sample standard deviation, to estimate  $\sigma$ , the population standard deviation. But in doing so, we cannot use formula (9–1). Because we do not know  $\sigma$ , we cannot use the  $z$  distribution. However, there is a remedy. We use the sample standard deviation and replace the  $z$  distribution with the  $t$  distribution.

The  $t$  distribution is a continuous probability distribution, with many similar characteristics to the  $z$  distribution. William Gosset, an English brewmaster, was the first to study the  $t$  distribution. He was particularly interested in the behavior of the distribution of the following statistic:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

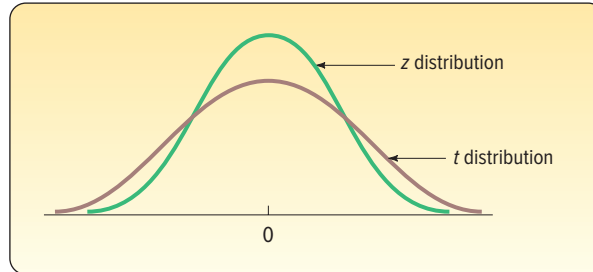
where  $s$  is an estimate of  $\sigma$ . He noticed differences between estimating  $\sigma$  based on  $s$ , especially when  $s$  was calculated from a very small sample. The  $t$  distribution and the standard normal distribution are shown graphically in Chart 9–1. Note particularly that the  $t$  distribution is flatter, more spread out, than the standard normal distribution. This is because the standard deviation of the  $t$  distribution is larger than that of the standard normal distribution.

The following characteristics of the  $t$  distribution are based on the assumption that the population of interest is normal, or nearly normal.

### STATISTICS IN ACTION

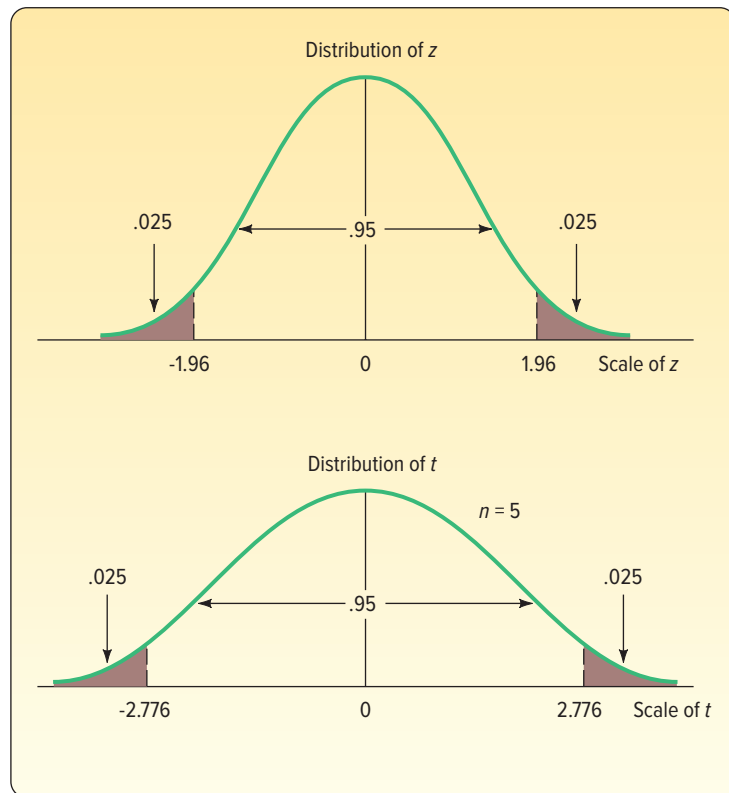
The  $t$  distribution was created by William Gosset, who was born in England in 1876 and died there in 1937. He worked for many years at Arthur Guinness, Sons and Company. In fact, in his later years he was in charge of the Guinness Brewery in London. Guinness preferred its employees to use pen names when publishing papers, so in 1908, when Gosset wrote "The Probable Error of a Mean," he used the name "Student." In this paper, he first described the properties of the  $t$  distribution and used it to monitor the brewing process so that the beer met Guinness's quality standards.

- It is, like the  $z$  distribution, a continuous distribution.
- It is, like the  $z$  distribution, bell-shaped and symmetrical.
- There is not one  $t$  distribution, but rather a family of  $t$  distributions. All  $t$  distributions have a mean of 0, but their standard deviations differ according to the sample size,  $n$ . There is a  $t$  distribution for a sample size of 20, another for a sample size of 22, and so on. The standard deviation for a  $t$  distribution with 5 observations is larger than for a  $t$  distribution with 20 observations.
- The  $t$  distribution is more spread out and flatter at the center than the standard normal distribution (see Chart 9–1). As the sample size increases, however, the  $t$  distribution approaches the standard normal distribution because the errors in using  $s$  to estimate  $\sigma$  decrease with larger samples.



**CHART 9–1** The Standard Normal Distribution and Student's  $t$  Distribution

Because Student's  $t$  distribution has a greater spread than the  $z$  distribution, the absolute value of  $t$  for a given level of confidence is greater in magnitude than the corresponding  $z$  value. Chart 9–2 shows the values of  $z$  for a 95% level of confidence and



**CHART 9–2** Values of  $z$  and  $t$  for the 95% Level of Confidence

of  $t$  for the same level of confidence when the sample size is  $n = 5$ . How we obtained the actual value of  $t$  will be explained shortly. For now, observe that for the same level of confidence the  $t$  distribution is flatter or more spread out than the standard normal distribution. Note that the margin of error for a 95% confidence interval using a  $t$ -statistic will be larger compared to using a  $z$ -statistic. The associated confidence interval using a  $t$ -statistic will be wider than an interval using a  $z$ -statistic.

To develop a confidence interval for the population mean using the  $t$ -distribution, we adjust formula (9–1) as follows.

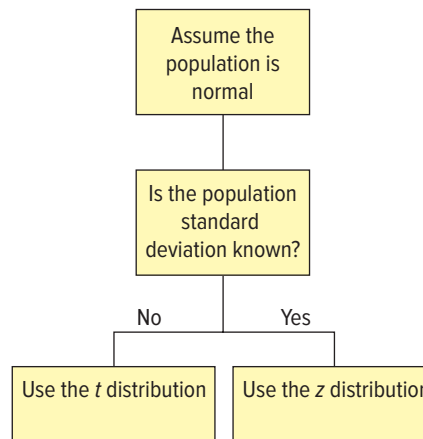
**CONFIDENCE INTERVAL FOR THE  
POPULATION MEAN,  $\sigma$  UNKNOWN**

$$\bar{x} \pm t \frac{s}{\sqrt{n}} \quad (9-2)$$

To determine a confidence interval for the population mean with an unknown population standard deviation, we:

1. Assume the sampled population is normal or approximately normal. This assumption may be questionable for small sample sizes, and it becomes more valid with larger sample sizes.
2. Estimate the population standard deviation ( $\sigma$ ) with the sample standard deviation ( $s$ ).
3. Use the  $t$  distribution rather than the  $z$  distribution.

We should be clear at this point. We base the decision to use the  $t$  or the  $z$  on whether or not we know  $\sigma$ , the population standard deviation. If we know the population standard deviation, then we use  $z$ . If we do not know the population standard deviation, then we must use  $t$ . Chart 9–3 summarizes the decision-making process.



**CHART 9–3** Determining When to Use the  $z$  Distribution or the  $t$  Distribution

The following example will illustrate a confidence interval for a population mean when the population standard deviation is unknown and how to find the appropriate value of  $t$  in a table.

**EXAMPLE**

A tire manufacturer wishes to investigate the tread life of its tires. A sample of 10 tires driven 50,000 miles revealed a sample mean of 0.32 inch of tread remaining with a standard deviation of 0.09 inch. Construct a 95% confidence interval

for the population mean. Would it be reasonable for the manufacturer to conclude that after 50,000 miles the population mean amount of tread remaining is 0.30 inch?

### SOLUTION

To begin, we assume the population distribution is normal. In this case, we don't have a lot of evidence, but the assumption is probably reasonable. We know the sample standard deviation is .09 inch. We use formula (9–2):

$$\bar{x} \pm t \frac{s}{\sqrt{n}}$$

From the information given,  $\bar{x} = 0.32$ ,  $s = 0.09$ , and  $n = 10$ . To find the value of  $t$ , we use Appendix B.5, a portion of which is reproduced in Table 9–2. The first step for locating  $t$  is to move across the columns identified for “Confidence Intervals” to the level of confidence requested. In this case, we want the 95% level of confidence, so we move to the column headed “95%.” The column on the left margin is identified as “*df*.” This refers to the number of degrees of freedom. The number of degrees of freedom is the number of observations in the sample minus the number of samples, written  $n - 1$ . In this case, it is  $10 - 1 = 9$ . Why did we decide there were 9 degrees of freedom? When sample statistics are being used, it is necessary to determine the number of values that are *free to vary*.

**TABLE 9–2** A Portion of the  $t$  Distribution

|           |   | Confidence Intervals |        |        |        |     |
|-----------|---|----------------------|--------|--------|--------|-----|
|           |   | 80%                  | 90%    | 95%    | 98%    | 99% |
| <i>df</i> | Level of Significance for One-Tailed Test |                      |        |        |        |     |
|           | 0.10                                      | 0.05                 | 0.025  | 0.010  | 0.005  |     |
|           | Level of Significance for Two-Tailed Test |                      |        |        |        |     |
|           | 0.20                                      | 0.10                 | 0.05   | 0.02   | 0.01   |     |
| 1         | 3.078                                     | 6.314                | 12.706 | 31.821 | 63.657 |     |
| 2         | 1.886                                     | 2.920                | 4.303  | 6.965  | 9.925  |     |
| 3         | 1.638                                     | 2.353                | 3.182  | 4.541  | 5.841  |     |
| 4         | 1.533                                     | 2.132                | 2.776  | 3.747  | 4.604  |     |
| 5         | 1.476                                     | 2.015                | 2.571  | 3.365  | 4.032  |     |
| 6         | 1.440                                     | 1.943                | 2.447  | 3.143  | 3.707  |     |
| 7         | 1.415                                     | 1.895                | 2.365  | 2.998  | 3.499  |     |
| 8         | 1.397                                     | 1.860                | 2.306  | 2.896  | 3.355  |     |
| 9         | 1.383                                     | 1.833                | 2.262  | 2.821  | 3.250  |     |
| 10        | 1.372                                     | 1.812                | 2.228  | 2.764  | 3.169  |     |

To illustrate the meaning of degrees of freedom: Assume that the mean of four numbers is known to be 5. The four numbers are 7, 4, 1, and 8. The deviations of these numbers from the mean must total 0. The deviations of +2, –1, –4, and +3 do total 0. If the deviations of +2, –1, and –4 are known, then the value of +3 is fixed (restricted) to satisfy the condition that the sum of the deviations must equal 0. Thus, 1 degree of freedom is lost in a sampling problem involving the



standard deviation of the sample because one number (the arithmetic mean) is known. For a 95% level of confidence and 9 degrees of freedom, we select the row with 9 degrees of freedom. The value of  $t$  is 2.262.

To determine the confidence interval, we substitute the values in formula (9-2).

$$\bar{x} \pm t \frac{s}{\sqrt{n}} = 0.32 \pm 2.262 \frac{0.09}{\sqrt{10}} = 0.32 \pm 0.64$$

The endpoints of the confidence interval are 0.256 and 0.384. How do we interpret this result? If we repeated this study 200 times, calculating the 95% confidence interval with each sample's mean and the standard deviation, we expect 190 of the intervals would include the population mean. Ten of the intervals would not include the population mean. This is the effect of sampling error. A further interpretation is to conclude that the population mean is in this interval. The manufacturer can be reasonably sure (95% confident) that the mean remaining tread depth is between 0.256 and 0.384 inch. Because the value of 0.30 is in this interval, it is possible that the mean of the population is 0.30.

Here is another example to clarify the use of confidence intervals. Suppose an article in your local newspaper reported that the mean time to sell a residential property in the area is 60 days. You select a random sample of 20 homes sold in the last year and find the mean selling time is 65 days. Based on the sample data, you develop a 95% confidence interval for the population mean. You find that the endpoints of the confidence interval are 62 days and 68 days. How do you interpret this result? You can be reasonably confident the population mean is within this range. The value proposed for the population mean, that is, 60 days, is not included in the interval. It is not likely that the population mean is 60 days. The evidence indicates the statement by the local newspaper may not be correct. To put it another way, it seems unreasonable to obtain the sample you did from a population that had a mean selling time of 60 days.

The following example will show additional details for determining and interpreting a confidence interval. We used Minitab to perform the calculations.

### ▶ EXAMPLE

The manager of the Inlet Square Mall, near Ft. Myers, Florida, wants to estimate the mean amount spent per shopping visit by customers. A sample of 20 customers reveals the following amounts spent.

|         |         |         |         |         |         |         |
|---------|---------|---------|---------|---------|---------|---------|
| \$48.16 | \$42.22 | \$46.82 | \$51.45 | \$23.78 | \$41.86 | \$54.86 |
| 37.92   | 52.64   | 48.59   | 50.82   | 46.94   | 61.83   | 61.69   |
| 49.17   | 61.46   | 51.35   | 52.68   | 58.84   | 43.88   |         |

What is the best estimate of the population mean? Determine a 95% confidence interval. Interpret the result. Would it be reasonable to conclude that the population mean is \$50? What about \$60?

### SOLUTION

The mall manager assumes that the population of the amounts spent follows the normal distribution. This is a reasonable assumption in this case. Additionally, the



©McGraw-Hill Education/Andrew Resek, photographer

confidence interval technique is quite powerful and tends to commit any errors on the conservative side if the population is not normal. We should not make the normality assumption when the population is severely skewed or when the distribution has “thick tails.” In this case, the normality assumption is reasonable.

The population standard deviation is not known. Hence, it is appropriate to use the  $t$  distribution and formula (9–2) to find the confidence interval. We use the Minitab system to find the mean and standard deviation of this sample. The results are shown below.

| Descriptive Statistics: Dollars Spent |        |       |         |                  |       |         |        |        |        |         |
|---------------------------------------|--------|-------|---------|------------------|-------|---------|--------|--------|--------|---------|
| Statistics                            |        |       |         |                  |       |         |        |        |        |         |
| Variable                              | N      | N*    | Mean    | SE Mean          | StDev | Minimum | Q1     | Median | Q3     | Maximum |
| Dollars Spent                         | 20     | 0     | 49.348  | 2.015            | 9.012 | 23.780  | 44.615 | 49.995 | 54.315 | 61.830  |
| 1-Sample t: Dollars Spent             |        |       |         |                  |       |         |        |        |        |         |
| Descriptive Statistics                |        |       |         |                  |       |         |        |        |        |         |
| N                                     | Mean   | StDev | SE Mean | 95% CI for $\mu$ |       |         |        |        |        |         |
| 20                                    | 49.348 | 9.012 | 2.015   | (45.130, 53.566) |       |         |        |        |        |         |
| $\mu$ : mean of Dollars Spent         |        |       |         |                  |       |         |        |        |        |         |

The mall manager does not know the population mean. The sample mean is the best estimate of that value. From the pictured Minitab output, the mean is \$49.348, which is the best estimate, the *point estimate*, of the unknown population mean.

We use formula (9–2) to find the confidence interval. The value of  $t$  is available from Appendix B.5. There are  $n - 1 = 20 - 1 = 19$  degrees of freedom. We move across the row with 19 degrees of freedom to the column for the 95% confidence level. The value at this intersection is 2.093. We substitute these values into formula (9–2) to find the confidence interval.

$$\bar{x} \pm t \frac{s}{\sqrt{n}} = \$49.348 \pm 2.093 \frac{\$9.012}{\sqrt{20}} = \$49.348 \pm \$4.218$$

The endpoints of the confidence interval are \$45.130 and \$53.566. It is reasonable to conclude that the population mean is in that interval.

The manager of Inlet Square wondered whether the population mean could have been \$50 or \$60. The value of \$50 is within the confidence interval. It is reasonable that the population mean could be \$50. The value of \$60 is not in the confidence interval. Hence, we conclude that the population mean is unlikely to be \$60.

Software to compute a confidence interval is also available in Excel. The output follows. Note that the sample mean (\$49.348) and the sample standard deviation (\$9.012) are the same as those in the Minitab calculations. Rather than directly computing a confidence interval, Excel only shows the margin of error, which is in the bottom line of the analysis (Confidence Level [95.0%]). It is the amount added and subtracted from the sample mean to form the endpoints of the confidence interval. This value is found from

$$t \frac{s}{\sqrt{n}} = 2.093 \frac{\$9.012}{\sqrt{20}} = \$4.218$$

| <i>Amount</i>           |        |
|-------------------------|--------|
| Mean                    | 49.35  |
| Standard Error          | 2.02   |
| Median                  | 50.00  |
| Mode                    | #N/A   |
| Standard Deviation      | 9.01   |
| Sample Variance         | 81.22  |
| Kurtosis                | 2.26   |
| Skewness                | -1.00  |
| Range                   | 38.05  |
| Minimum                 | 23.78  |
| Maximum                 | 61.83  |
| Sum                     | 986.96 |
| Count                   | 20.00  |
| Confidence Level(95.0%) | 4.22   |

Before doing the confidence interval exercises, we would like to point out a useful characteristic of the *t* distribution that will allow us to use the *t* table to quickly find both *z* and *t* values. Earlier in this section, on page 253, we detailed the characteristics of the *t* distribution. The last point indicated that as we increase the sample size, the *t* distribution approaches the *z* distribution. In fact, when we reach an infinitely large sample, the *t* distribution is exactly equal to the *z* distribution.

To explain, Table 9–3 is a portion of Appendix B.5, with the degrees of freedom between 4 and 99 omitted. To find the appropriate *z* value for a 95% confidence interval,

**TABLE 9–3** Student’s *t* Distribution

| <i>df</i><br>(degrees of freedom) | Confidence Interval                                 |       |        |        |        |         |
|-----------------------------------|---|-------|--------|--------|--------|---------|
|                                   | 80%   | 90%   | 95%    | 98%    | 99%    | 99.9%   |
|                                   | Level of Significance for One-Tailed Test, $\alpha$ |       |        |        |        |         |
|                                   | 0.1   | 0.05  | 0.025  | 0.01   | 0.005  | 0.0005  |
|                                   | Level of Significance for Two-Tailed Test, $\alpha$ |       |        |        |        |         |
|                                   | 0.2   | 0.1   | 0.05   | 0.02   | 0.01   | 0.001   |
| 1                                 | 3.078   | 6.314 | 12.706 | 31.821 | 63.657 | 636.619 |
| 2                                 | 1.886   | 2.920 | 4.303  | 6.965  | 9.925  | 31.599  |
| 3                                 | 1.638   | 2.353 | 3.182  | 4.541  | 5.841  | 12.924  |
| ⋮                                 | ⋮   | ⋮     | ⋮      | ⋮      | ⋮      | ⋮       |
| 100                               | 1.290   | 1.660 | 1.984  | 2.364  | 2.626  | 3.390   |
| 120                               | 1.289   | 1.658 | 1.980  | 2.358  | 2.617  | 3.373   |
| 140                               | 1.288   | 1.656 | 1.977  | 2.353  | 2.611  | 3.361   |
| 160                               | 1.287   | 1.654 | 1.975  | 2.350  | 2.607  | 3.352   |
| 180                               | 1.286   | 1.653 | 1.973  | 2.347  | 2.603  | 3.345   |
| 200                               | 1.286   | 1.653 | 1.972  | 2.345  | 2.601  | 3.340   |
| ∞                                 | 1.282   | 1.645 | 1.960  | 2.326  | 2.576  | 3.291   |

we begin by going to the confidence interval section and selecting the column headed “95%.” Move down that column to the last row, which is labeled “∞,” or infinite degrees of freedom. The value reported is 1.960, the same value that we found using the standard normal distribution in Appendix B.3. This confirms the convergence of the  $t$  distribution to the  $z$  distribution.

What does this mean for us? Instead of searching in the body of the  $z$  table, we can go to the last row of the  $t$  table and find the appropriate value to build a confidence interval. An additional benefit is that the values have three decimal places. So, using this table for a 90% confidence interval, go down the column headed “90%” and see the value 1.645, which is a more precise  $z$  value that can be used for the 90% confidence level. Other  $z$  values for 98% and 99% confidence intervals are also available with three decimals. Note that we will use the  $t$  table, which is summarized in Table 9–3, to find the  $z$  values with three decimals for all following exercises and problems.

## SELF-REVIEW 9–2



Dottie Kleman is the “Cookie Lady.” She bakes and sells cookies at locations in the Philadelphia area. Ms. Kleman is concerned about absenteeism among her workers. The information below reports the number of days absent for a sample of 10 workers during the last two-week pay period.

4   1   2   2   1   2   2   1   0   3

- Determine the mean and the standard deviation of the sample.
- What is the population mean? What is the best estimate of that value?
- Develop a 95% confidence interval for the population mean. Assume that the population distribution is normal.
- Explain why the  $t$  distribution is used as a part of the confidence interval.
- Is it reasonable to conclude that the typical worker does not miss any days during a pay period?

## EXERCISES

- Use Appendix B.5 to locate the value of  $t$  under the following conditions.
  - The sample size is 12 and the level of confidence is 95%.
  - The sample size is 20 and the level of confidence is 90%.
  - The sample size is 8 and the level of confidence is 99%.
- Use Appendix B.5 to locate the value of  $t$  under the following conditions.
  - The sample size is 15 and the level of confidence is 95%.
  - The sample size is 24 and the level of confidence is 98%.
  - The sample size is 12 and the level of confidence is 90%.
- The owner of Britten’s Egg Farm wants to estimate the mean number of eggs produced per chicken. A sample of 20 chickens shows they produced an average of 20 eggs per month with a standard deviation of 2 eggs per month.
  - What is the value of the population mean? What is the best estimate of this value?
  - Explain why we need to use the  $t$  distribution. What assumption do you need to make?
  - For a 95% confidence interval, what is the value of  $t$ ?
  - Develop the 95% confidence interval for the population mean.
  - Would it be reasonable to conclude that the population mean is 21 eggs? What about 25 eggs?
- The U.S. Dairy Industry wants to estimate the mean yearly milk consumption. A sample of 16 people reveals the mean yearly consumption to be 45 gallons

with a standard deviation of 20 gallons. Assume the population distribution is normal.

- a. What is the value of the population mean? What is the best estimate of this value?
  - b. Explain why we need to use the  $t$  distribution. What assumption do you need to make?
  - c. For a 90% confidence interval, what is the value of  $t$ ?
  - d. Develop the 90% confidence interval for the population mean.
  - e. Would it be reasonable to conclude that the population mean is 48 gallons?
13. **FILE** Merrill Lynch Securities and Health Care Retirement Inc. are two large employers in downtown Toledo, Ohio. They are considering jointly offering child care for their employees. As a part of the feasibility study, they wish to estimate the mean weekly child care cost of their employees. A sample of 10 employees who use child care reveals the following amounts spent last week.

\$107   \$92   \$97   \$95   \$105   \$101   \$91   \$99   \$95   \$104

Develop a 90% confidence interval for the population mean. Interpret the result.

14. **FILE** The Buffalo, New York, Chamber of Commerce wants to estimate the mean time workers who are employed in the downtown area spend getting to work. A sample of 15 workers reveals the following number of minutes spent traveling.

14   24   24   19   24   7   31   20  
26   23   23   28   16   15   21

Develop a 98% confidence interval for the population mean. Interpret the result.

### LO9-3

Compute and interpret a confidence interval for a population proportion.

## A CONFIDENCE INTERVAL FOR A POPULATION PROPORTION

The material presented so far in this chapter uses the ratio scale of measurement. That is, we use such variables as incomes, weights, distances, and ages. We now want to consider situations such as the following:

- The career services director at Southern Technical Institute reports that 80% of its graduates enter the job market in a position related to their field of study.
- A company representative claims that 45% of Burger King sales are made at the drive-through window.
- A survey of homes in the Chicago area indicated that 85% of the new construction had central air conditioning.
- A recent survey of married men between the ages of 35 and 50 found that 63% felt that both partners should earn a living.

These examples illustrate the nominal scale of measurement when the outcome is limited to two values. In these cases, an observation is classified into one of two mutually exclusive groups. For example, a graduate of Southern Tech either entered the job market in a position related to his or her field of study or not. A particular Burger King customer either made a purchase at the drive-through window or did not make a purchase at the drive-through window. We can talk about the groups in terms of **proportions**.

**PROPORTION** The fraction, ratio, or percent indicating the part of the sample or the population having a particular trait of interest.



As an example of a proportion, a recent survey indicated that 92 out of 100 people surveyed favored the continued use of daylight saving time in the summer. The sample proportion is  $92/100$ , or  $.92$ , or  $92\%$ . If we let  $p$  represent the sample proportion,  $x$  the number of “successes,” and  $n$  the number of items sampled, we can determine a sample proportion as follows.

**SAMPLE PROPORTION**

$$p = \frac{x}{n}$$

**(9–3)**

The population proportion is identified by  $\pi$ . Therefore,  $\pi$  refers to the percent of successes in the population. Recall from Chapter 6 that  $\pi$  is the proportion of “successes” in a binomial distribution. This continues our practice of using Greek letters to identify population parameters and Roman letters to identify sample statistics.

To develop a confidence interval for a proportion, we need to meet two requirements:

1. The binomial conditions, discussed in Chapter 6, have been met. These conditions are:
  - a. The sample data are the number of successes in  $n$  trials.
  - b. There are only two possible outcomes. (We usually label one of the outcomes a “success” and the other a “failure.”)
  - c. The probability of a success remains the same from one trial to the next.
  - d. The trials are independent. This means the outcome on one trial does not affect the outcome on another.
2. The values  $n\pi$  and  $n(1 - \pi)$  should both be greater than or equal to 5. This allows us to invoke the central limit theorem and employ the standard normal distribution, that is,  $z$ , to complete a confidence interval.

Developing a point estimate for a population proportion and a confidence interval for a population proportion is similar to doing so for a mean. To illustrate, John Gail is running for Congress from the Third District of Nebraska. From a random sample of 100 voters in the district, 60 indicate they plan to vote for him in the upcoming election. The sample proportion is  $.60$ , but the population proportion is unknown. That is, we do not know what proportion of voters in the *population* will vote for Mr. Gail. The sample value,  $.60$ , is the best estimate we have of the unknown population parameter. So we let  $p$ , which is  $.60$ , be an estimate of  $\pi$ , which is not known.

To develop a confidence interval for a population proportion, we use:

**CONFIDENCE INTERVAL FOR A POPULATION PROPORTION**

$$p \pm z \sqrt{\frac{p(1-p)}{n}}$$

**(9–4)**

An example will help to explain the details of determining a confidence interval and interpreting the result.

**EXAMPLE**

The union representing the Bottle Blowers of America (BBA) is considering a proposal to merge with the Teamsters Union. According to BBA union bylaws, at least three-fourths of the union membership must approve any merger. A random sample of 2,000 current BBA members reveals 1,600 plan to vote for the merger proposal. What is the estimate of the population proportion? Develop a 95% confidence interval for the population proportion. Basing your decision on this sample information, can you conclude that the necessary proportion of BBA members favor the merger? Why?

**SOLUTION**

First, calculate the sample proportion from formula (9–3). It is .80, found by

$$p = \frac{x}{n} = \frac{1,600}{2,000} = .80$$

Thus, we estimate that 80% of the population favor the merger proposal. We determine the 95% confidence interval using formula (9–4). The  $z$  value corresponding to the 95% level of confidence is 1.96.

$$p \pm z \sqrt{\frac{p(1-p)}{n}} = .80 \pm 1.96 \sqrt{\frac{.80(1-.80)}{2,000}} = .80 \pm .018$$

The endpoints of the confidence interval are .782 and .818. The lower endpoint is greater than .75. Hence, we conclude that the merger proposal will likely pass because the interval estimate includes only values greater than 75% of the union membership.

**STATISTICS IN ACTION**

The results of many surveys include confidence intervals. For example, a recent survey of 800 TV viewers in Toledo, Ohio, found 44% watched the evening news on the local CBS affiliate. The article also reported a margin of error of 3.4%. The margin of error is actually the amount that is added and subtracted from the point estimate to find the endpoints of a confidence interval. For a 95% level of confidence, the margin of error is:

$$\begin{aligned} z \sqrt{\frac{p(1-p)}{n}} \\ &= 1.96 \sqrt{\frac{.44(1-.44)}{800}} \\ &= 0.034 \end{aligned}$$

The estimate of the proportion of all TV viewers in Toledo, Ohio, who watch the local news on CBS is between  $(.44 - .034)$  and  $(.44 + .034)$ , or 40.6% and 47.4%.

To review the interpretation of the confidence interval: If the poll was conducted 100 times with 100 different samples, we expect the confidence intervals constructed from 95 of the samples would contain the true population proportion. In addition, the interpretation of a confidence interval can be very useful in decision making and play a very important role, especially on election night. For example, Cliff Obermeyer is running for Congress from the Sixth District of New Jersey. Suppose 500 voters are contacted upon leaving the polls and 275 indicate they voted for Mr. Obermeyer. We will assume that the exit poll of 500 voters is a random sample of those voting in the Sixth District. That means that 55% of those in the sample voted for Mr. Obermeyer. Based on formula (9–3):

$$p = \frac{x}{n} = \frac{275}{500} = .55$$

Now, to be assured of election, he must earn *more than* 50% of the votes in the population of those voting. At this point, we know a point estimate, which is .55, of the population of voters that will vote for him. But we do not know the percent in the population that will ultimately vote for the candidate. So the question is: Could we take a sample of 500 voters from a population where 50% or less of the voters support Mr. Obermeyer and find that 55% of the sample support him? To put it another way, could the sampling error, which is  $p - \pi = .55 - .50 = .05$  be due to chance, or is the population of voters who support Mr. Obermeyer greater than .50? If we develop a confidence interval for the sample proportion and find that the lower endpoint is greater than .50, then we conclude that the proportion of voters supporting Mr. Obermeyer is greater than .50. What does that mean? Well, it means he should be elected! What if .50 is in the interval? Then we conclude that he is not assured of a majority and we cannot conclude he will be elected. In this case, using the 95% significance level and formula (9–4):

$$p \pm z \sqrt{\frac{p(1-p)}{n}} = .55 \pm 1.96 \sqrt{\frac{.55(1-.55)}{500}} = .55 \pm .044$$

The endpoints of the confidence interval are  $.55 - .044 = .506$  and  $.55 + .044 = .594$ . The value of .50 is not in this interval. So we conclude that probably *more than 50%* of the voters support Mr. Obermeyer and that is enough to get him elected.

Is this procedure ever used? Yes! It is exactly the procedure used by polling organizations, television networks, and surveys of public opinion on election night.

## SELF-REVIEW 9-3



A market research consultant was hired to estimate the proportion of homemakers who associate the brand name of a laundry detergent with the container's shape and color. The consultant randomly selected 1,400 homemakers. From the sample, 420 were able to identify the brand by name based only on the shape and color of the container.

- Estimate the value of the population proportion.
- Develop a 99% confidence interval for the population proportion.
- Interpret your findings.

## EXERCISES

- The owner of the West End Kwick Fill Gas Station wishes to determine the proportion of customers who pay at the pump using a credit card or debit card. He surveys 100 customers and finds that 80 paid at the pump.
  - Estimate the value of the population proportion.
  - Develop a 95% confidence interval for the population proportion.
  - Interpret your findings.
- Ms. Maria Wilson is considering running for mayor of Bear Gulch, Montana. Before completing the petitions, she decides to conduct a survey of voters in Bear Gulch. A sample of 400 voters reveals that 300 would support her in the November election.
  - Estimate the value of the population proportion.
  - Develop a 99% confidence interval for the population proportion.
  - Interpret your findings.
- The Fox TV network is considering replacing one of its prime-time crime investigation shows with a new family-oriented comedy show. Before a final decision is made, network executives designed an experiment to estimate the proportion of their viewers who would prefer the comedy show over the crime investigation show. A random sample of 400 viewers was selected and asked to watch the new comedy show and the crime investigation show. After viewing the shows, 250 indicated they would watch the new comedy show and suggested it replace the crime investigation show.
  - Estimate the value of the population proportion of people who would prefer the comedy show.
  - Develop a 99% confidence interval for the population proportion of people who would prefer the comedy show.
  - Interpret your findings.
- Schadek Silkscreen Printing Inc. purchases plastic cups and imprints them with logos for sporting events, proms, birthdays, and other special occasions. Zack Schadek, the owner, received a large shipment this morning. To ensure the quality of the shipment, he selected a random sample of 300 cups and inspected them for defects. He found 15 to be defective.
  - What is the estimated proportion defective in the population?
  - Develop a 95% confidence interval for the proportion defective.
  - Zack has an agreement with his supplier that if 10% or more of the cups are defective, he can return the order. Should he return this lot? Explain your decision.

### LO9-4

Calculate the required sample size to estimate a population proportion or population mean.

## CHOOSING AN APPROPRIATE SAMPLE SIZE

When working with confidence intervals, one important variable is sample size. However, in practice, sample size is not a variable. It is a decision we make so that our estimate of a population parameter is a good one. Our decision is based on three factors:

- The margin of error the researcher will tolerate.
- The level of confidence desired, for example, 95%.
- The variation or dispersion of the population being studied.



The first factor is the *margin of error*. It is designated as  $E$  and is the amount that is added and subtracted to the sample mean (or sample proportion) to determine the endpoints of the confidence interval. For example, in a study of wages, we decide to estimate the mean wage of the population with a margin of error of plus or minus \$1,000. Or, in an opinion poll, we may decide that we want to estimate the population proportion with a margin of error of plus or minus 3.5%. The margin of error is the amount of error we are willing to tolerate in estimating a population parameter. You may wonder why we do not choose small margins of error. There is a trade-off between the margin of error and sample size. A small margin of error will require a larger sample and more money and time to collect the sample. A larger margin of error will permit a smaller sample and result in a wider confidence interval.

The second factor is the *level of confidence*. In working with confidence intervals, we logically choose relatively high levels of confidence such as 95% and 99%. To compute the sample size, we need the  $z$ -statistic that corresponds to the chosen level of confidence. The 95% level of confidence corresponds to a  $z$  value of 1.96, and a 90% level of confidence corresponds to a  $z$  value of 1.645 (using the  $t$  table). Notice that larger sample sizes (and more time and money to collect the sample) correspond with higher levels of confidence. Also, notice that we use a  $z$ -statistic.

The third factor to determine the sample size is the *population standard deviation*. If the population is widely dispersed, a large sample is required to get a good estimate. On the other hand, if the population is concentrated (homogeneous), the required sample size to get a good estimate will be smaller. Often, we do not know the population standard deviation. Here are three suggestions to estimate the population standard deviation.

1. **Conduct a pilot study.** This is the most common method. Suppose we want an estimate of the number of hours per week worked by students enrolled in the College of Business at the University of Texas. To test the validity of our questionnaire, we use it on a small sample of students. From this small sample, we compute the standard deviation of the number of hours worked and use this value as the population standard deviation.
2. **Use a comparable study.** Use this approach when there is an estimate of the standard deviation from another study. Suppose we want to estimate the number of hours worked per week by refuse workers. Information from certain state or federal agencies that regularly study the workforce may provide a reliable value to use for the population standard deviation.
3. **Use a range-based approach.** To use this approach, we need to know or have an estimate of the largest and smallest values in the population. Recall from Chapter 3, the Empirical Rule states that virtually all the observations could be expected to be within plus or minus 3 standard deviations of the mean, assuming that the distribution follows the normal distribution. Thus, the distance between the largest and the smallest values is 6 standard deviations. We can estimate the standard deviation as one-sixth of the range. For example, the director of operations at University Bank wants to estimate the number of ATM transactions per month made by college students. She believes that the distribution of ATM transactions follows the normal distribution. The minimum and maximum of ATM transactions per month are 2 and 50, so the range is 48, found by  $(50 - 2)$ . Then the estimated value of the population standard deviation would be eight ATM transactions per month,  $48/6$ .

## Sample Size to Estimate a Population Mean

To estimate a population mean, we can express the interaction among these three factors and the sample size in the following formula. Notice that this formula is the margin of error used to calculate the endpoints of confidence intervals to estimate a population mean! See formula (9–1).

$$E = z \frac{\sigma}{\sqrt{n}}$$

Solving this equation for  $n$  yields the following result.

**SAMPLE SIZE FOR ESTIMATING  
THE POPULATION MEAN**

$$n = \left( \frac{z\sigma}{E} \right)^2 \quad (9-5)$$

where:

$n$  is the size of the sample.

$z$  is the standard normal  $z$  value corresponding to the desired level of confidence.

$\sigma$  is the population standard deviation.

$E$  is the maximum allowable error.

The result of this calculation is not always a whole number. When the outcome is not a whole number, the usual practice is to round up *any* fractional result to the next whole number. For example, 201.21 would be rounded up to 202.

**EXAMPLE**

A student in public administration wants to estimate the mean monthly earnings of city council members in large cities. She can tolerate a margin of error of \$100 in estimating the mean. She would also prefer to report the interval estimate with a 95% level of confidence. The student found a report by the Department of Labor that reported a standard deviation of \$1,000. What is the required sample size?

**SOLUTION**

The maximum allowable error,  $E$ , is \$100. The value of  $z$  for a 95% level of confidence is 1.96, and the value of the standard deviation is \$1,000. Substituting these values into formula (9-5) gives the required sample size as:

$$n = \left( \frac{z\sigma}{E} \right)^2 = \left( \frac{(1.96)(\$1,000)}{\$100} \right)^2 = (19.6)^2 = 384.16$$

The computed value of 384.16 is rounded up to 385. A sample of 385 is required to meet the specifications. If the student wants to increase the level of confidence, for example to 99%, this will require a larger sample. Using the  $t$  table with infinite degrees of freedom, the  $z$  value for a 99% level of confidence is 2.576.

$$n = \left( \frac{z\sigma}{E} \right)^2 = \left( \frac{(2.576)(\$1,000)}{\$100} \right)^2 = (25.76)^2 = 663.58$$

We recommend a sample of 664. Observe how much the change in the confidence level changed the size of the sample. An increase from the 95% to the 99% level of confidence resulted in an increase of 279 observations, or 72%  $[(664/385) \times 100]$ . This would greatly increase the cost of the study, in terms of both time and money. Hence, the level of confidence should be considered carefully.

## Sample Size to Estimate a Population Proportion

To determine the sample size to estimate a population proportion, the same three factors need to be specified:

1. The margin of error.
2. The desired level of confidence.
3. The variation or dispersion of the population being studied.

For the binomial distribution, the margin of error is:

$$E = z \sqrt{\frac{\pi(1 - \pi)}{n}}$$

Solving this equation for  $n$  yields the following equation

**SAMPLE SIZE FOR THE  
POPULATION PROPORTION**

$$n = \pi(1 - \pi) \left( \frac{z}{E} \right)^2 \quad (9-6)$$

where:

$n$  is the size of the sample.

$z$  is the standard normal  $z$  value corresponding to the desired level of confidence.

$\pi$  is the population proportion.

$E$  is the maximum allowable error.

As before, the  $z$  value is associated with our choice of confidence level. We also decide the margin of error,  $E$ . However, the population variance of the binomial distribution is represented by  $\pi(1 - \pi)$ . To estimate the population variance, we need a value of the population proportion. If a reliable value cannot be determined with a pilot study or found in a comparable study, then a value of .50 can be used for  $\pi$ . Note that  $\pi(1 - \pi)$  has the largest value using 0.50 and, therefore, without a good estimate of the population proportion, using 0.50 as an estimate of  $\pi$  overstates the sample size. Using a larger sample size will not hurt the estimate of the population proportion.

► **EXAMPLE**

The student in the previous example also wants to estimate the proportion of cities that have private refuse collectors. The student wants to estimate the population proportion with a margin of error of .10, prefers a level of confidence of 90%, and has no estimate for the population proportion. What is the required sample size?

**SOLUTION**

The estimate of the population proportion is to be within .10, so  $E = .10$ . The desired level of confidence is .90, which corresponds to a  $z$  value of 1.645, using the  $t$  table with infinite degrees of freedom. Because no estimate of the population proportion is available, we use .50. The suggested number of observations is

$$n = (.5)(1 - .5) \left( \frac{1.645}{.10} \right)^2 = 67.65$$

The student needs a random sample of 68 cities.

**SELF-REVIEW 9-4**



A university's office of research wants to estimate the arithmetic mean grade point average (GPA) of all graduating seniors during the past 10 years. GPAs range between 2.0 and 4.0. The estimate of the population mean GPA should be within plus or minus .05 of the population mean. Based on prior experience, the population standard deviation is 0.279. Using a 99% level of confidence, how many student records need to be selected?

## EXERCISES

19. A population's standard deviation is 10. We want to estimate the population mean within 2, with a 95% level of confidence. How large a sample is required?
20. We want to estimate the population mean within 5, with a 99% level of confidence. The population standard deviation is estimated to be 15. How large a sample is required?
21. The estimate of the population proportion should be within plus or minus .05, with a 95% level of confidence. The best estimate of the population proportion is .15. How large a sample is required?
22. The estimate of the population proportion should be within plus or minus .10, with a 99% level of confidence. The best estimate of the population proportion is .45. How large a sample is required?
23. A large on-demand video streaming company is designing a large-scale survey to determine the mean amount of time corporate executives watch on-demand television. A small pilot survey of 10 executives indicated that the mean time per week is 12 hours, with a standard deviation of 3 hours. The estimate of the mean viewing time should be within one-quarter hour. The 95% level of confidence is to be used. How many executives should be surveyed?
24. A processor of carrots cuts the green top off each carrot, washes the carrots, and inserts six to a package. Twenty packages are inserted in a box for shipment. Each box of carrots should weigh 20.4 pounds. The processor knows that the standard deviation of box weight is 0.5 pound. The processor wants to know if the current packing process meets the 20.4 weight standard. How many boxes must the processor sample to be 95% confident that the estimate of the population mean is within 0.2 pound?
25. Suppose the U.S. president wants to estimate the proportion of the population that supports his current policy toward revisions in the health care system. The president wants the estimate to be within .04 of the true proportion. Assume a 95% level of confidence. The president's political advisors found a similar survey from two years ago that reported that 60% of people supported health care revisions.
  - a. How large of a sample is required?
  - b. How large of a sample would be necessary if no estimate were available for the proportion supporting current policy?
26. Past surveys reveal that 30% of tourists going to Las Vegas to gamble spend more than \$1,000. The Visitor's Bureau of Las Vegas wants to update this percentage.
  - a. How many tourists should be randomly selected to estimate the population proportion with a 90% confidence level and a 1% margin of error?
  - b. The Bureau feels the sample size determined above is too large. What can be done to reduce the sample? Based on your suggestion, recalculate the sample size.

## CHAPTER SUMMARY

- I. A point estimate is a single value (statistic) used to estimate a population value (parameter).
- II. A confidence interval is a range of values within which the population parameter is expected to occur.
  - A. The factors that determine the width of a confidence interval for a mean are:
    1. The number of observations in the sample,  $n$ .
    2. The variability in the population, usually estimated by the sample standard deviation,  $s$ .
    3. The level of confidence.
      - a. To determine the confidence limits when the population standard deviation is known, we use the  $z$  distribution. The formula is

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}} \quad (9-1)$$

- b. To determine the confidence limits when the population standard deviation is unknown, we use the  $t$  distribution. The formula is

$$\bar{x} \pm t \frac{S}{\sqrt{n}} \quad (9-2)$$

- III. The major characteristics of the  $t$  distribution are:

- A. It is a continuous distribution.
  - B. It is mound-shaped and symmetrical.
  - C. It is flatter, or more spread out, than the standard normal distribution.
  - D. There is a family of  $t$  distributions, depending on the number of degrees of freedom.
- IV. A proportion is a ratio, fraction, or percent that indicates the part of the sample or population that has a particular characteristic.
- A. A sample proportion,  $p$ , is found by  $x$ , the number of successes, divided by  $n$ , the number of observations.
  - B. We construct a confidence interval for a sample proportion from the following formula.

$$p \pm z \sqrt{\frac{p(1-p)}{n}} \quad (9-4)$$

- V. We can determine an appropriate sample size for estimating both means and proportions.

- A. There are three factors that determine the sample size when we wish to estimate the mean.
  1. The margin of error,  $E$ .
  2. The desired level of confidence.
  3. The variation in the population.

The formula to determine the sample size for the mean is

$$n = \left( \frac{z\sigma}{E} \right)^2 \quad (9-5)$$

- B. There are three factors that determine the sample size when we wish to estimate a proportion.
  1. The margin of error,  $E$ .
  2. The desired level of confidence.
  3. A value for  $\pi$  to calculate the variation in the population.

The formula to determine the sample size for a proportion is

$$n = \pi(1-\pi) \left( \frac{z}{E} \right)^2 \quad (9-6)$$

## CHAPTER EXERCISES

27. A random sample of 85 group leaders, supervisors, and similar personnel at General Motors revealed that, on average, they spent 6.5 years in a particular job before being promoted. The standard deviation of the sample was 1.7 years. Construct a 95% confidence interval.
28. A state meat inspector in Iowa has been given the assignment of estimating the mean net weight of packages of ground chuck labeled "3 pounds." Of course, he realizes that the weights cannot always be precisely 3 pounds. A sample of 36 packages reveals the mean weight to be 3.01 pounds, with a standard deviation of 0.03 pound.
- a. What is the estimated population mean?
  - b. Determine a 95% confidence interval for the population mean.
29. As part of their business promotional package, the Milwaukee Chamber of Commerce would like an estimate of the mean cost per month to lease a one-bedroom apartment. The mean cost per month for a random sample of 40 apartments currently available for lease was \$884. The standard deviation of the sample was \$50.
- a. Develop a 98% confidence interval for the population mean.
  - b. Would it be reasonable to conclude that the population mean is \$950 per month?

- 30.** A recent survey of 50 executives who were laid off during a recent recession revealed it took a mean of 26 weeks for them to find another position. The standard deviation of the sample was 6.2 weeks. Construct a 95% confidence interval for the population mean. Is it reasonable that the population mean is 28 weeks? Justify your answer.
- 31.** Marty Rowatti recently assumed the position of director of the YMCA of South Jersey. He would like some data on how long current members of the YMCA have been members. To investigate, suppose he selects a random sample of 40 current members. The mean length of membership for the sample is 8.32 years and the standard deviation is 3.07 years.
- What is the mean of the population?
  - Develop a 90% confidence interval for the population mean.
  - The previous director, in the summary report she prepared as she retired, indicated the mean length of membership was now “almost 10 years.” Does the sample information substantiate this claim? Cite evidence.
- 32.** The American Restaurant Association collected information on the number of meals eaten outside the home per week by young married couples. A survey of 60 couples showed the sample mean number of meals eaten outside the home was 2.76 meals per week, with a standard deviation of 0.75 meal per week. Construct a 99% confidence interval for the population mean.
- 33.** The National Collegiate Athletic Association (NCAA) reported that college football assistant coaches spend a mean of 70 hours per week on coaching and recruiting during the season. A random sample of 50 assistant coaches showed the sample mean to be 68.6 hours, with a standard deviation of 8.2 hours.
- Using the sample data, construct a 99% confidence interval for the population mean.
  - Does the 99% confidence interval include the value suggested by the NCAA? Interpret this result.
  - Suppose you decided to switch from a 99% to a 95% confidence interval. Without performing any calculations, will the interval increase, decrease, or stay the same? Which of the values in the formula will change?
- 34.** The human relations department of Electronics Inc. would like to include a dental plan as part of the benefits package. The question is: How much does a typical employee and his or her family spend per year on dental expenses? A sample of 45 employees reveals the mean amount spent last year was \$1,820, with a standard deviation of \$660.
- Construct a 95% confidence interval for the population mean.
  - The information from part (a) was given to the president of Electronics Inc. He indicated he could afford \$1,700 of dental expenses per employee. Is it possible that the population mean could be \$1,700? Justify your answer.
- 35.** A student conducted a study and reported that the 95% confidence interval for the mean ranged from 46 to 54. He was sure that the mean of the sample was 50, that the standard deviation of the sample was 16, and that the sample size was at least 30, but he could not remember the exact number. Can you help him out?
- 36.** A recent study by the American Automobile Dealers Association surveyed a random sample of 20 dealers. The data revealed a mean amount of profit per car sold was \$290, with a standard deviation of \$125. Develop a 95% confidence interval for the population mean of profit per car.
- 37.** A study of 25 graduates of four-year public colleges revealed the mean amount owed by a student in student loans was \$55,051. The standard deviation of the sample was \$7,568. Construct a 90% confidence interval for the population mean. Is it reasonable to conclude that the mean of the population is actually \$55,000? Explain why or why not.
- 38.** An important factor in selling a residential property is the number of times real estate agents show a home. A sample of 15 homes recently sold in the Buffalo, New York, area revealed the mean number of times a home was shown was 24 and the standard deviation of the sample was 5 people. Develop a 98% confidence interval for the population mean.
- 39.** **FILE** In 2003, the Accreditation Council for Graduate Medical Education (ACGME) implemented new rules limiting work hours for all residents. A key component of these rules is

that residents should work no more than 80 hours per week. The following is the number of weekly hours worked in 2017 by a sample of residents at the Tidelands Medical Center.

84 86 84 86 79 82 87 81 84 78 74 86

- What is the point estimate of the population mean for the number of weekly hours worked at the Tidelands Medical Center?
  - Develop a 90% confidence interval for the population mean.
  - Is the Tidelands Medical Center within the ACGME guideline? Why?
40. **FILE** PrintTech Inc. is introducing a new line of inkjet printers and would like to promote the number of pages a user can expect from a print cartridge. A sample of 10 cartridges revealed the following number of pages printed.

2,698 2,028 2,474 2,395 2,372 2,475 1,927 3,006 2,334 2,379

- What is the point estimate of the population mean?
  - Develop a 95% confidence interval for the population mean.
41. **FILE** Dr. Susan Benner is an industrial psychologist. She is currently studying stress among executives of Internet companies. She has developed a questionnaire that she believes measures stress. A score above 80 indicates stress at a dangerous level. A random sample of 15 executives revealed the following stress level scores.

94 78 83 90 78 99 97 90 97 90 93 94 100 75 84

- Find the mean stress level for this sample. What is the point estimate of the population mean?
  - Construct a 95% confidence level for the population mean.
  - According to Dr. Benner's test, is it reasonable to conclude that the mean stress level of Internet executives is 80? Explain.
42. Pharmaceutical companies promote their prescription drugs using television advertising. In a survey of 80 randomly sampled television viewers, 10 indicated that they asked their physician about using a prescription drug they saw advertised on TV. Develop a 95% confidence interval for the proportion of viewers who discussed a drug seen on TV with their physician. Is it reasonable to conclude that 25% of the viewers discuss an advertised drug with their physician?
43. HighTech Inc. randomly tests its employees about company policies. Last year, in the 400 random tests conducted, 14 employees failed the test. Develop a 99% confidence interval for the proportion of applicants that fail the test. Would it be reasonable to conclude that 5% of the employees cannot pass the company policy test? Explain.
44. During a national debate on changes to health care, a cable news service performs an opinion poll of 500 small-business owners. It shows that 65% of small-business owners do not approve of the changes. Develop a 95% confidence interval for the proportion opposing health care changes. Comment on the result.
45. There are 20,000 eligible voters in York County, South Carolina. A random sample of 500 York County voters revealed 350 plan to vote to return Louella Miller to the state senate. Construct a 99% confidence interval for the proportion of voters in the county who plan to vote for Ms. Miller. From this sample information, is it reasonable to conclude that Ms. Miller will receive a majority of the votes?
46. In a poll to estimate presidential popularity, each person in a random sample of 1,000 voters was asked to agree with one of the following statements:
- The president is doing a good job.
  - The president is doing a poor job.
  - I have no opinion.
- A total of 560 respondents selected the first statement, indicating they thought the president was doing a good job.
- Construct a 95% confidence interval for the proportion of respondents who feel the president is doing a good job.
  - Based on your interval in part (a), is it reasonable to conclude that a majority of the population believes the president is doing a good job?

- 47.** It is estimated that 60% of U.S. households subscribe to cable TV. You would like to verify this statement for your class in mass communications. If you want your estimate to be within 5 percentage points, with a 95% level of confidence, how many households should you sample?
- 48.** You wish to estimate the mean number of travel days per year for salespeople. The mean of a small pilot study was 150 days, with a standard deviation of 14 days. If you want to estimate the population mean within 2 days, how many salespeople should you sample? Use the 90% confidence level.
- 49.** You want to estimate the mean family income in a rural area of central Indiana. The question is, how many families should be sampled? In a pilot sample of 10 families, the standard deviation of the sample was \$500. The sponsor of the survey wants you to use the 95% confidence level. The estimate is to be within \$100. How many families should be interviewed?
- 50.** *Families USA*, a monthly magazine that discusses issues related to health and health costs, surveyed 20 of its subscribers. It found that the annual health insurance premiums for a family with coverage through an employer averaged \$10,979. The standard deviation of the sample was \$1,000.
- Based on this sample information, develop a 90% confidence interval for the population mean yearly premium.
  - How large a sample is needed to find the population mean within \$250 at 99% confidence?
- 51.** Passenger comfort is influenced by the amount of pressurization in an airline cabin. Higher pressurization permits a closer-to-normal environment and a more relaxed flight. A study by an airline user group recorded the equivalent air pressure on 30 randomly chosen flights. The study revealed a mean equivalent air pressure of 8,000 feet with a standard deviation of 300 feet.
- Develop a 99% confidence interval for the population mean equivalent air pressure.
  - How large a sample is needed to find the population mean within 25 feet at 95% confidence?
- 52.** A survey of 25 randomly sampled judges employed by the state of Florida found that they earned an average wage (including benefits) of \$65.00 per hour. The sample standard deviation was \$6.25 per hour.
- What is the population mean? What is the best estimate of the population mean?
  - Develop a 99% confidence interval for the population mean wage (including benefits) for these employees.
  - How large a sample is needed to assess the population mean with an allowable error of \$1.00 at 95% confidence?
- 53.** Based on a sample of 50 U.S. citizens, the American Film Institute found that a typical American spent 78 hours watching movies last year. The standard deviation of this sample was 9 hours.
- Develop a 95% confidence interval for the population mean number of hours spent watching movies last year.
  - How large a sample should be used to be 90% confident the sample mean is within 1.0 hour of the population mean?
- 54.** Dylan Jones kept careful records of the fuel efficiency of his new car. After the first nine times he filled up the tank, he found the mean was 23.4 miles per gallon (mpg) with a sample standard deviation of 0.9 mpg.
- Compute the 95% confidence interval for his mpg.
  - How many times should he fill his gas tank to obtain a margin of error below 0.1 mpg?
- 55.** A survey of 36 randomly selected iPhone owners showed that the purchase price has a mean of \$650 with a sample standard deviation of \$24.
- Compute the standard error of the sample mean.
  - Compute the 95% confidence interval for the mean.
  - How large a sample is needed to estimate the population mean within \$10?
- 56.** You plan to conduct a survey to find what proportion of the workforce has two or more jobs. You decide on the 95% confidence level and a margin of error of 2%. A pilot survey reveals that 5 of the 50 sampled hold two or more jobs. How many in the workforce should be interviewed to meet your requirements?



57. A study conducted several years ago reported that 21 percent of public accountants changed companies within 3 years. The American Institute of CPAs would like to update the study. They would like to estimate the population proportion of public accountants who changed companies within 3 years with a margin of error of 3% and a 95% level of confidence.
- To update this study, the files of how many public accountants should be studied?
  - How many public accountants should be contacted if no previous estimates of the population proportion are available?
58. As part of an annual review of its accounts, a discount brokerage selected a random sample of 36 customers and reviewed the value of their accounts. The mean was \$32,000 with a sample standard deviation of \$8,200. What is a 90% confidence interval for the mean account value of the population of customers?
59. The National Weight Control Registry tries to mine secrets of success from people who lost at least 30 pounds and kept it off for at least a year. It reports that out of 2,700 registrants, 459 were on a low-carbohydrate diet (less than 90 grams a day).
- Develop a 95% confidence interval for the proportion of people on a low-carbohydrate diet.
  - Is it possible that the population percentage is 18%?
  - How large a sample is needed to estimate the proportion within 0.5%?
60. Near the time of an election, a cable news service performs an opinion poll of 1,000 probable voters. It shows that the Republican contender has an advantage of 52% to 48%.
- Develop a 95% confidence interval for the proportion favoring the Republican candidate.
  - Estimate the probability that the Democratic candidate is actually leading.
  - Repeat the above analysis based on a sample of 3,000 probable voters.
61. A sample of 352 subscribers to *Wired* magazine shows the mean time spent using the Internet is 13.4 hours per week, with a sample standard deviation of 6.8 hours. Find the 95% confidence interval for the mean time *Wired* subscribers spend on the Internet.
62. The Tennessee Tourism Institute (TTI) plans to sample information center visitors entering the state to learn the fraction of visitors who plan to camp in the state. Current estimates are that 35% of visitors are campers. How many visitors would you sample to estimate the population proportion of campers with a 95% confidence level and an allowable error of 2%?

## DATA ANALYTICS

(The data for these exercises are available in Connect.)

63. **FILE** Refer to the North Valley Real Estate data, which report information on homes sold in the area during the last year. Select a random sample of 20 homes.
- Based on your random sample of 20 homes, develop a 95% confidence interval for the mean selling price of the homes.
  - Based on your random sample of 20 homes, develop a 95% confidence interval for the mean days on the market.
  - Based on your random sample of 20 homes, develop a 95% confidence interval for the proportion of homes with a pool.
  - Suppose that North Valley Real Estate employs several agents. Each agent will be randomly assigned 20 homes to sell. The agents are highly motivated to sell homes based on the commissions they earn. They are also concerned about the 20 homes they are assigned to sell. Using the confidence intervals you created, write a general memo informing the agents about the characteristics of the homes they may be assigned to sell.
  - What would you do if your confidence intervals did not include the mean of all 105 homes? How could this happen?
64. **FILE** Refer to the Baseball 2016 data, which report information on the 30 Major League Baseball teams for the 2016 season. Assume the 2016 data represent a sample.
- Develop a 95% confidence interval for the mean number of home runs per team.
  - Develop a 95% confidence interval for the mean batting average by each team.
  - Develop a 95% confidence interval for the mean earned run average (ERA) for each team.
65. **FILE** Refer to the Lincolnville School District bus data.
- Develop a 95% confidence interval for the mean bus maintenance cost.
  - Develop a 95% confidence interval for the mean bus odometer miles.
  - Write a business memo to the state transportation official to report your results.

**PRACTICE TEST****Part 1—Objective**

1. A \_\_\_\_\_ is a single value computed from sample information used to estimate a population parameter.
2. A \_\_\_\_\_ is a range of values within which the population parameter is likely to occur.
3. Assuming the same sample size and the same standard deviation, a 90% confidence interval will be \_\_\_\_\_ a 95% confidence interval. (equal to, wider than, narrower than, can't tell)
4. A \_\_\_\_\_ shows the fraction of a sample that has a particular characteristic.
5. For a 95% level of confidence, approximately \_\_\_\_\_ percent of the similarly constructed intervals will include the population parameter being estimated.
6. To construct a confidence interval for a mean, the  $z$  distribution is used only when the \_\_\_\_\_ is known. (population mean, population standard deviation, sample size, population size)
7. To develop a confidence interval for a proportion, the four conditions of what probability distribution must be met? \_\_\_\_\_ (normal, Poisson,  $t$  distribution, binomial)
8. As the degrees of freedom increase, the  $t$  distribution \_\_\_\_\_. (approaches the binomial distribution, exceeds the normal distribution, approaches the  $z$  distribution, becomes more positively skewed)
9. The \_\_\_\_\_ has no effect on the size of the sample. (level of confidence, margin of error, population median, variability in the population)
10. To locate the appropriate  $t$  value, which is not necessary? (degrees of freedom, level of confidence, population mean)

**Part 2—Problems**

1. A recent study of 26 Conway, South Carolina, residents revealed they had lived at their current address for a mean of 9.3 years, with a sample standard deviation of 2 years.
  - a. What is the population mean?
  - b. What is the best estimate of the population mean?
  - c. What is the standard error of the mean?
  - d. Develop a 90% confidence interval for the population mean.
2. A recent federal report indicated 27% of children ages 2 to 5 ate vegetables at least five times a week. How large a sample is necessary to estimate the true population proportion within 2% with a 98% level of confidence? Be sure to use the evidence in the federal report.
3. The Philadelphia Regional Transport Authority wishes to estimate the proportion of central city workers that use public transportation to get to work. A recent study reported that of 100 workers, 64 used public transportation. Develop a 95% confidence interval.

# 10

# One-Sample Tests of Hypothesis



©paulista/Shutterstock

- ▲ **DOLE PINEAPPLE INC.** is concerned that the 16-ounce can of sliced pineapple is being overfilled. Assume the standard deviation of the process is .03 ounce. The quality control department took a random sample of 50 cans and found that the arithmetic mean weight was 16.05 ounces. At the 5% level of significance, can we conclude that the mean weight is greater than 16 ounces? Determine the  $p$ -value. (See Exercise 24 and **LO10-4**.)

## LEARNING OBJECTIVES

---

When you have completed this chapter, you will be able to:

- LO10-1** Explain the process of testing a hypothesis.
- LO10-2** Apply the six-step procedure for testing a hypothesis.
- LO10-3** Distinguish between a one-tailed and a two-tailed test of hypothesis.
- LO10-4** Conduct a test of a hypothesis about a population mean.
- LO10-5** Compute and interpret a  $p$ -value.
- LO10-6** Use a  $t$  statistic to test a hypothesis.

## INTRODUCTION

Chapter 8 began our study of sampling and statistical inference. We described how we could select a random sample to estimate the value of a population parameter. For example, we selected a sample of five employees at Spence Sprockets, found the number of years of service for each sampled employee, computed the mean years of service, and used the sample mean to estimate the mean years of service for all employees. In other words, we estimated a population parameter from a sample statistic.

Chapter 9 continued the study of statistical inference by developing a confidence interval. A confidence interval is a range of values within which we expect the population parameter to occur. In this chapter, rather than develop a range of values within which we expect the population parameter to occur, we develop a procedure to test the validity of a statement about a population parameter. Some examples of statements we might want to test are:

- The mean speed of automobiles passing milepost 150 on the West Virginia Turnpike is 68 miles per hour.
- The mean number of miles driven by those leasing a Chevy Trail-Blazer for 3 years is 32,000 miles.
- The mean time an American family lives in a particular single-family dwelling is 11.8 years.
- In 2016, the mean starting salary for a graduate from a four-year business program is \$51,541.
- According to the Kelley Blue Book ([www.kbb.com](http://www.kbb.com)), a 2017 Ford Edge averages 21 miles per gallon in the city.
- The mean cost to remodel a kitchen is \$20,000.



©Russell Illig/Getty Images

This chapter and several of the following chapters cover statistical hypothesis testing. We begin by defining what we mean by a statistical hypothesis and statistical hypothesis testing. Next, we outline the steps in statistical hypothesis testing. Then, we conduct tests of hypothesis for means. In the last section of the chapter, we describe possible errors due to sampling in hypothesis testing.

### LO10-1

Explain the process of testing a hypothesis.

## WHAT IS HYPOTHESIS TESTING?

The terms *hypothesis testing* and *testing a hypothesis* are used interchangeably. Hypothesis testing starts with a statement, or assumption, about a population parameter—such as the population mean. This statement is referred to as a hypothesis.

**HYPOTHESIS** A statement about a population parameter subject to verification.

A hypothesis might be that the mean monthly commission of sales associates in retail electronics stores, such as Conn's Homeplus, is \$2,000. We cannot contact all Conn sales associates to determine that the mean is \$2,000. The cost of locating and interviewing every Conn electronics sales associate in the United States would be exorbitant. To test the validity of the hypothesis ( $\mu = \$2,000$ ), we must select a sample from the population of all Conn electronics sales associates, calculate sample statistics, and,

based on certain decision rules, reject or fail to reject the hypothesis. A sample mean of \$1,000 per month is much less than \$2,000 per month, and we would most likely reject the hypothesis. However, suppose the sample mean is \$1,995. Can we attribute the \$5 difference between \$1,995 and \$2,000 to sampling error? Or is this difference of \$5 statistically significant?

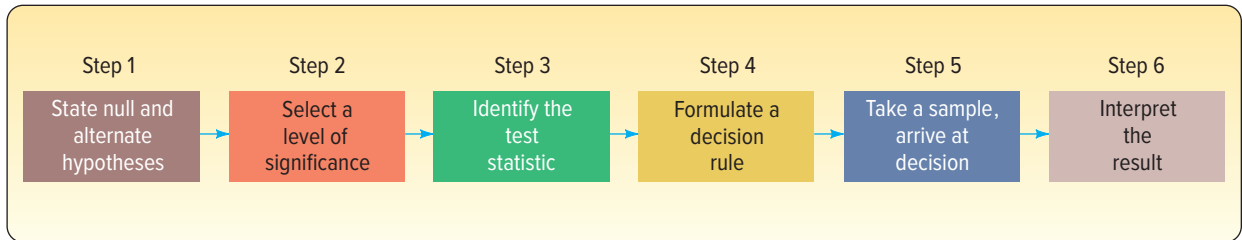
**HYPOTHESIS TESTING** A procedure based on sample evidence and probability theory to determine whether the hypothesis is a reasonable statement.

**LO10-2**

Apply the six-step procedure for testing a hypothesis.

## SIX-STEP PROCEDURE FOR TESTING A HYPOTHESIS

There is a six-step procedure that systematizes hypothesis testing; when we get to step 6, we are ready to interpret the results of the test based on the decision to reject or not reject the hypothesis. However, hypothesis testing as used by statisticians does not provide proof that something is true, in the manner in which a mathematician “proves” a statement. It does provide a kind of “proof beyond a reasonable doubt,” in the manner of the court system. Hence, there are specific rules of evidence, or procedures, that are followed. The steps are shown in the following diagram. We will discuss in detail each of the steps.



### Step 1: State the Null Hypothesis ( $H_0$ ) and the Alternate Hypothesis ( $H_1$ )

The first step is to state the hypothesis being tested. It is called the **null hypothesis**, designated  $H_0$ , and read “*H sub zero*.” The capital letter  $H$  stands for hypothesis, and the subscript zero implies “no difference.” There is usually a “not” or a “no” term in the null hypothesis, meaning that there is “no change.” For example, the null hypothesis is that the mean number of miles driven on the steel-belted tire is not different from 60,000. The null hypothesis would be written  $H_0: \mu = 60,000$ . Generally speaking, the null hypothesis is developed for the purpose of testing. We either reject or fail to reject the null hypothesis. The null hypothesis is a statement that is not rejected unless our sample data provide convincing evidence that it is false.

We should emphasize that if the null hypothesis is not rejected on the basis of the sample data, we cannot say that the null hypothesis is true. To put it another way, failing to reject the null hypothesis does not prove that  $H_0$  is true; it means we have *failed to disprove*  $H_0$ . To prove without any doubt the null hypothesis is true, the population parameter would have to be known. To actually determine it, we would have to test, survey, or count every item in the population. This is usually not feasible. The alternative is to take a sample from the population.

Often, the null hypothesis begins by stating, “There is no *significant* difference between . . .” or, “The mean impact strength of the glass is not *significantly* different

from . . .” When we select a sample from a population, the sample statistic is usually numerically different from the hypothesized population parameter. As an illustration, suppose the hypothesized impact strength of a glass plate is 70 psi, and the mean impact strength of a sample of 12 glass plates is 69.5 psi. We must make a decision about the difference of 0.5 psi. Is it a true difference, that is, a significant difference, or is the difference between the sample statistic (69.5) and the hypothesized population parameter (70.0) due to chance (sampling error)? To answer this question, we conduct a test of significance, commonly referred to as a test of hypothesis. To define what is meant by a null hypothesis:

**NULL HYPOTHESIS** A statement about the value of a population parameter developed for the purpose of testing numerical evidence.

The **alternate hypothesis** describes what you will conclude if you reject the null hypothesis. It is written  $H_1$  and is read “*H sub one*.” It is also referred to as the research hypothesis. The alternate hypothesis is accepted if the sample data provide us with enough statistical evidence that the null hypothesis is false.

**ALTERNATE HYPOTHESIS** A statement that is accepted if the sample data provide sufficient evidence that the null hypothesis is false.

The following example will help clarify what is meant by the null hypothesis and the alternate hypothesis. A recent article indicated the mean age of U.S. commercial aircraft is 15 years. To conduct a statistical test regarding this statement, the first step is to determine the null and the alternate hypotheses. The null hypothesis represents the current or reported condition. It is written  $H_0: \mu = 15$ . The alternate hypothesis is that the statement is not true, that is,  $H_1: \mu \neq 15$ . It is important to remember that no matter how the problem is stated, *the null hypothesis will always contain the equal sign*. The equal sign (=) will never appear in the alternate hypothesis. Why? Because the null hypothesis is the statement being tested, and we need a specific value to include in our calculations. We turn to the alternate hypothesis only if the data suggest the null hypothesis is untrue.

## Step 2: Select a Level of Significance

After setting up the null hypothesis and alternate hypothesis, the next step is to state the **level of significance**.

The level of significance is designated  $\alpha$ , the Greek letter alpha. It is also sometimes called the level of risk. This may be a more appropriate term because it is the risk you take of rejecting the null hypothesis when it is really true.

**LEVEL OF SIGNIFICANCE** The probability of rejecting the null hypothesis when it is true.

There is no one level of significance that is applied to all tests. A decision is made to use the .05 level (often stated as the 5% level), the .01 level, or the .10 level. As a probability, the significance level must be between 0 and 1. However, it is reasonable to choose probabilities that are small. Traditionally, the .05 level is selected for consumer research projects, .01 for quality assurance, and .10 for political polling. You, the researcher, must decide on the level of significance *before* formulating a decision rule and collecting sample data.



©tcsaba/Shutterstock

To illustrate how it is possible to reject a true hypothesis, suppose a firm manufacturing personal computers uses a large number of printed circuit boards. Suppliers bid on the boards, and the one with the lowest bid is awarded a sizable contract. Suppose the contract specifies that the computer manufacturer's quality assurance department will randomly sample all incoming shipments of circuit boards. If more than 6% of the boards sampled are substandard, the shipment will be rejected. The null hypothesis is that the incoming shipment of boards meets the quality standards of the contract and contains 6% or less defective boards. The alternate hypothesis is that more than 6% of the boards are defective.

A shipment of 4,000 circuit boards was received from Allied Electronics, and the quality assurance department selected a random sample of 50 circuit boards for testing. Of the 50 circuit boards sampled, 4 boards, or 8%, were substandard. The shipment was rejected because it exceeded the maximum of 6% substandard printed circuit boards. If the shipment was actually substandard, then the decision to return the boards to the supplier was correct.

However, because of sampling error, there is a small probability of an incorrect decision. Suppose there were only 40, or 4%, defective boards in the shipment (well under the 6% threshold) and 4 of these 40 were randomly selected in the sample of 50. The sample evidence indicates that the percentage of defective boards is 8% (4 out of 50 is 8%), so we reject the shipment. But, in fact, of the 4,000 boards, there are only 40 defective units. The true defect rate is 1.00%. In this instance, our sample evidence estimates 8% defective but there is only 1% defective in the population. Based on the sample evidence, an incorrect decision was made. In terms of hypothesis testing, we rejected the null hypothesis when we should have failed to reject the null hypothesis. By rejecting a true null hypothesis, we committed a Type I error. The probability of committing a Type I error is represented by the Greek letter alpha ( $\alpha$ ).

**TYPE I ERROR** Rejecting the null hypothesis,  $H_0$ , when it is true.

The other possible error in hypothesis testing is called a Type II error. The probability of committing a Type II error is designated by the Greek letter beta ( $\beta$ ).

**TYPE II ERROR** Not rejecting the null hypothesis when it is false.

The firm manufacturing personal computers would commit a Type II error if, unknown to the manufacturer, an incoming shipment of printed circuit boards from Allied Electronics contained 15% substandard boards, yet the shipment was accepted. How could this happen? A random sample of 50 boards could have 2 (4%) substandard boards, and 48 good boards. According to the stated procedure, because the sample contained less than 6% substandard boards, the decision is to accept the shipment. This is a Type II error. While this event is extremely unlikely, it is possible based on the process of randomly sampling from a population.

In retrospect, the researcher cannot study every item or individual in the population. Thus, there is a possibility of two types of error—a Type I error, wherein the null hypothesis is rejected when it should not be rejected, and a Type II error, wherein the null hypothesis is not rejected when it should have been rejected.

We often refer to the probability of these two possible errors as *alpha*,  $\alpha$ , and *beta*,  $\beta$ . Alpha ( $\alpha$ ) is the probability of making a Type I error, and beta ( $\beta$ ) is the probability of making a Type II error. The following table summarizes the decisions a researcher could make and the possible consequences.

| Null Hypothesis | Researcher            |                  |
|-----------------|-----------------------|------------------|
|                 | Does Not Reject $H_0$ | Rejects $H_0$    |
| $H_0$ is true   | Correct decision      | Type I error     |
| $H_0$ is false  | Type II error         | Correct decision |

### Step 3: Select the Test Statistic

There are many **test statistics**. In this chapter, we use both  $z$  and  $t$  as the test statistics. In later chapters, we will use such test statistics as  $F$  and  $\chi^2$ , called chi-square.

**TEST STATISTIC** A value, determined from sample information, used to decide whether to reject the null hypothesis.

In hypothesis testing for the mean ( $\mu$ ) when  $\sigma$  is known, the test statistic  $z$  is computed by:

**TESTING A MEAN,  $\sigma$  KNOWN**

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

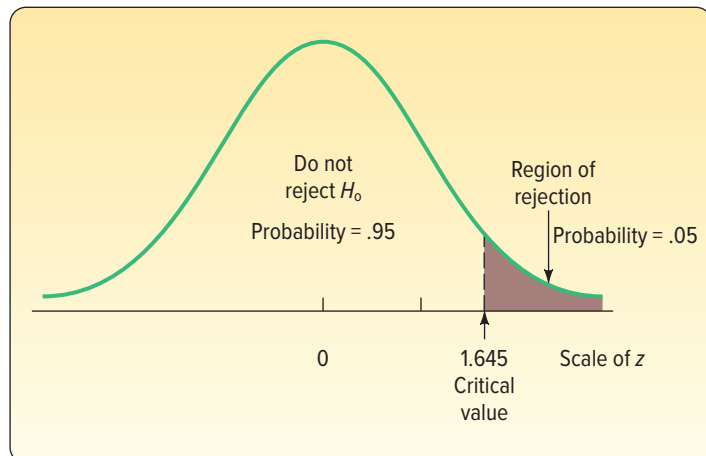
**(10–1)**

The  $z$  value is based on the sampling distribution of  $\bar{x}$ , which follows the normal distribution with a mean ( $\mu_{\bar{x}}$ ) equal to  $\mu$  and a standard deviation  $\sigma_{\bar{x}}$ , which is equal to  $\sigma/\sqrt{n}$ . We can thus determine whether the difference between  $\bar{x}$  and  $\mu$  is statistically significant by finding the number of standard deviations  $\bar{x}$  is from  $\mu$ , using formula (10–1).

### Step 4: Formulate the Decision Rule

A decision rule is a statement of the specific conditions under which the null hypothesis is rejected and the conditions under which it is not rejected. The region or area of rejection defines the location of all those values that are so large or so small that the probability of their occurrence under a true null hypothesis is rather remote.

Chart 10–1 portrays the rejection region for a test of significance that will be conducted later in the chapter.



**CHART 10–1** Sampling Distribution of the Statistic  $z$ , a Right-Tailed Test, .05 Level of Significance



**STATISTICS IN ACTION**

During World War II, allied military planners needed estimates of the number of German tanks. The information provided by traditional spying methods was not reliable, but statistical methods proved to be valuable. For example, espionage and reconnaissance led analysts to estimate that 1,550 tanks were produced during June 1941. However, using the serial numbers of captured tanks and statistical analysis, military planners estimated that only 244 tanks were produced. The actual number produced, as determined from German production records, was 271. The estimate using statistical analysis turned out to be much more accurate. A similar type of analysis was used to estimate the number of Iraqi tanks destroyed during the Persian Gulf War in 1991.

Note in the chart that:

- The area where the null hypothesis is not rejected is to the left of 1.645. We will explain how to get the 1.645 value shortly.
- The area of rejection is to the right of 1.645.
- A one-tailed test is being applied. (This will also be explained later.)
- The .05 level of significance was chosen.
- The sampling distribution of the statistic  $z$  follows the normal probability distribution.
- The value 1.645 separates the regions where the null hypothesis is rejected and where it is not rejected.
- The value 1.645 is the **critical value**.

**CRITICAL VALUE** The dividing point between the region where the null hypothesis is rejected and the region where it is not rejected.

## Step 5: Make a Decision

The fifth step in hypothesis testing is to compute the value of the test statistic, compare its value to the critical value, and make a decision to reject or not to reject the null hypothesis. Referring to Chart 10–1, if, based on sample information,  $z$  is computed to be 2.34, the null hypothesis is rejected at the .05 level of significance. The decision to reject  $H_0$  was made because 2.34 lies in the region of rejection, that is, beyond 1.645. We reject the null hypothesis, reasoning that it is highly improbable that a computed  $z$  value this large is due to sampling error (chance).

Had the computed value been 1.645 or less, say 0.71, the null hypothesis is not rejected. It is reasoned that such a small computed value could be attributed to chance, that is, sampling error. As we have emphasized, only one of two decisions is possible in hypothesis testing—either reject or do not reject the null hypothesis.

However, because the decision is based on a sample, it is always possible to make either of two decision errors. It is possible to make a Type I error when the null hypothesis is rejected when it should not be rejected. Or it is also possible to make a Type II error when the null hypothesis is not rejected and it should have been rejected. Fortunately, we select the probability of making a Type I error,  $\alpha$  (alpha), and we can compute the probabilities associated with a Type II error,  $\beta$  (beta).

## Step 6: Interpret the Result

The final step in the hypothesis testing procedure is to interpret the results. The process does not end with the value of a sample statistic or the decision to reject or not reject the null hypothesis. What can we say or report based on the results of the statistical test? Here are two examples:

- An investigative reporter for a Colorado newspaper reports that the mean monthly income of convenience stores in the state is \$130,000. You decide to conduct a test of hypothesis to verify the report. The null hypothesis and the alternate hypothesis are:

$$H_0: \mu = \$130,000$$

$$H_1: \mu \neq \$130,000$$

A sample of convenience stores provides a sample mean and standard deviation, and you compute a  $z$ -statistic. The results of the hypothesis test result in a decision to not reject the null hypothesis. How do you interpret the result? Be cautious with

**STATISTICS IN ACTION**

LASIK is a 15-minute surgical procedure that uses a laser to reshape an eye's cornea with the goal of improving eyesight. Research shows that about 5% of all surgeries involve complications such as glare, corneal haze, overcorrection or undercorrection of vision, and loss of vision. In a statistical sense, the research tests a null hypothesis that the surgery will not improve eyesight with the alternative hypothesis that the surgery will improve eyesight. The sample data of LASIK surgery shows that 5% of all cases result in complications. The 5% represents a Type I error rate. When a person decides to have the surgery, he or she expects to reject the null hypothesis. In 5% of future cases, this expectation will not be met. (Source: *American Academy of Ophthalmology Journal*, Vol. 16, no. 43.)

your interpretation because by not rejecting the null hypothesis, you did not prove the null hypothesis to be true. Based on the sample data, the difference between the sample mean and hypothesized population mean was not large enough to reject the null hypothesis.

- In a recent speech to students, the dean of the College of Business reported that the mean credit card debt for college students is \$3,000. You decide to conduct a test of the dean's statement, or hypothesis, to investigate the statement's truth. The null hypothesis and the alternate hypothesis are:

$$H_0: \mu = \$3,000$$

$$H_1: \mu \neq \$3,000$$

A random sample of college students provides a sample mean and standard deviation, and you compute a z-statistic. The hypothesis test results in a decision to reject the null hypothesis. How do you interpret the result? The sample evidence does not support the dean's statement. Based on the sample data, the mean amount of student credit card debt is different from \$3,000. You have disproved the null hypothesis with a stated probability of a Type I error,  $\alpha$ . That is, there is a small probability that the decision to reject the null hypothesis was an error due to random sampling.

**SUMMARY OF THE STEPS IN HYPOTHESIS TESTING**

1. Establish the null hypothesis ( $H_0$ ) and the alternate hypothesis ( $H_1$ ).
2. Select the level of significance, that is,  $\alpha$ .
3. Select an appropriate test statistic.
4. Formulate a decision rule based on steps 1, 2, and 3 above.
5. Make a decision regarding the null hypothesis based on the sample information.
6. Interpret the results of the test.

Before actually conducting a test of hypothesis, we describe the difference between a one-tailed and a two-tailed hypothesis test.

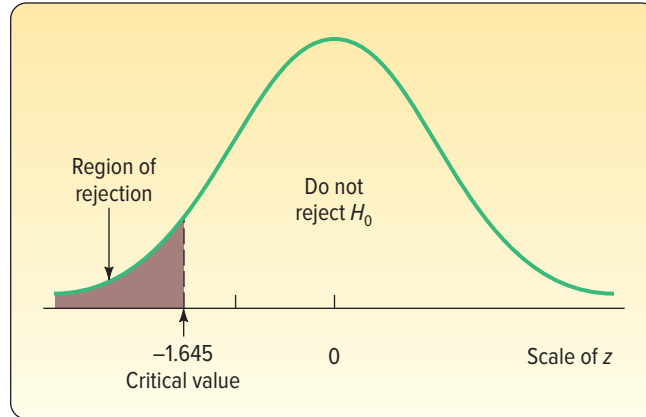
**LO10-3**

Distinguish between a one-tailed and a two-tailed test of hypothesis.

**ONE-TAILED AND TWO-TAILED HYPOTHESIS TESTS**

Refer to Chart 10–1. It shows a one-tailed test. It is called a one-tailed test because the rejection region is only in one tail of the curve. In this case, it is in the right, or upper, tail of the curve. To illustrate, suppose that the packaging department at General Foods Corporation is concerned that some boxes of Grape Nuts are significantly overweight. The cereal is packaged in 453-gram boxes, so the null hypothesis is  $H_0: \mu \leq 453$ . This is read, “the population mean ( $\mu$ ) is equal to or less than 453.” The alternate hypothesis is, therefore,  $H_1: \mu > 453$ . This is read, “ $\mu$  is greater than 453.” Note that the inequality sign in the alternate hypothesis ( $>$ ) points to the region of rejection in the upper tail. (See Chart 10–1.) Also observe that the null hypothesis includes the equal sign. That is,  $H_0: \mu \leq 453$ . The equality condition always appears in  $H_0$ , never in  $H_1$ .

Chart 10–2 portrays a situation where the rejection region is in the left (lower) tail of the standard normal distribution. As an illustration, consider the problem of automobile manufacturers, large automobile leasing companies, and other organizations that purchase large quantities of tires. They want the tires to average, say, 60,000 miles of wear under normal usage. They will, therefore, reject a shipment of tires if tests reveal that the mean life of the tires is significantly below 60,000 miles. They gladly accept a shipment if the mean life is greater than 60,000 miles! They are not concerned with this possibility, however. They are concerned only if they have sample evidence to conclude



**CHART 10–2** Sampling Distribution for the Statistic  $z$ , Left-Tailed Test, .05 Level of Significance

that the tires will average less than 60,000 miles of useful life. Thus, the test is set up to satisfy the concern of the automobile manufacturers that *the mean life of the tires is not less than 60,000 miles*. This statement appears in the null hypothesis. The null and alternate hypotheses in this case are written  $H_0: \mu \geq 60,000$  and  $H_1: \mu < 60,000$ .

One way to determine the location of the rejection region is to look at the direction in which the inequality sign in the alternate hypothesis is pointing (either  $<$  or  $>$ ). In the tire wear problem, it is pointing to the left, and the rejection region is therefore in the left tail.

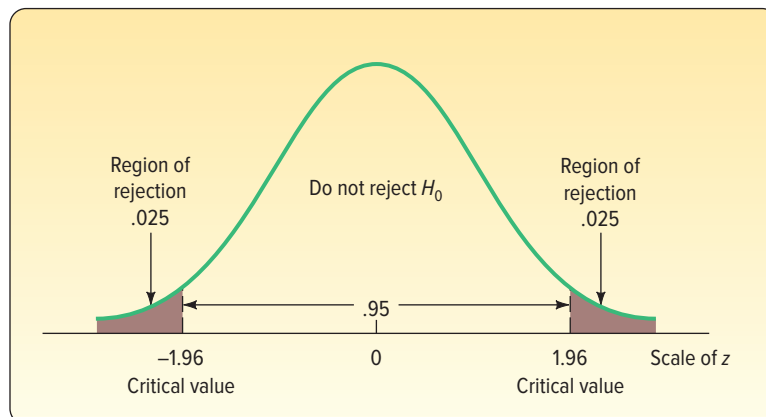
In summary, a test is *one-tailed* when the alternate hypothesis,  $H_1$ , states a direction, such as:

- $H_0$ : The mean income of female stockbrokers is *less than or equal to* \$65,000 per year.
- $H_1$ : The mean income of female stockbrokers is *greater than* \$65,000 per year.

If no direction is specified in the alternate hypothesis, we use a *two-tailed* test. Changing the previous problem to illustrate, we can say:

- $H_0$ : The mean income of female stockbrokers is \$65,000 per year.
- $H_1$ : The mean income of female stockbrokers is *not equal to* \$65,000 per year.

If the null hypothesis is rejected and  $H_1$  accepted in the two-tailed case, the mean income could be significantly greater than \$65,000 per year or it could be significantly less than \$65,000 per year. To accommodate these two possibilities, the 5% area of rejection is divided equally into the two tails of the sampling distribution (2.5% each). Chart 10–3 shows the two areas and the critical values. Note that the total area in the normal distribution is 1.0000, found by  $.9500 + .0250 + .0250$ .



**CHART 10–3** Regions of Nonrejection and Rejection for a Two-Tailed Test, .05 Level of Significance

**LO10-4**

Conduct a test of a hypothesis about a population mean.

## HYPOTHESIS TESTING FOR A POPULATION MEAN: KNOWN POPULATION STANDARD DEVIATION

### A Two-Tailed Test

An example will show the details of the six-step hypothesis testing procedure. We also wish to use a two-tailed test. That is, we are *not* concerned whether the sample results are larger or smaller than the proposed population mean. Rather, we are interested in whether it is *different from* the proposed value for the population mean. We begin, as we did in the previous chapter, with a situation in which we have historical information about the population and in fact know its standard deviation.

#### EXAMPLE

Jamestown Steel Company manufactures and assembles desks and other office equipment at several plants in western New York state. The weekly production of the Model A325 desk at the Fredonia plant follows a normal probability distribution with a mean of 200 and a standard deviation of 16. Recently, because of market expansion, new production methods have been introduced and new employees hired. The vice president of manufacturing would like to investigate whether there has been a *change* in the weekly production of the Model A325 desk. Is the mean number of desks produced at the Fredonia plant *different from* 200 at the .01 significance level?



©Robert Nicholas/Getty Images

#### SOLUTION

In this example, we know two important pieces of information: (1) the population of weekly production follows the normal distribution, and (2) the standard deviation of this normal distribution is 16 desks per week. So it is appropriate to use the z-statistic. We use the statistical hypothesis testing procedure to investigate whether the production rate has changed from 200 per week.

**Step 1: State the null hypothesis and the alternate hypothesis.** The null hypothesis is “The population mean is 200.” The alternate hypothesis is “The mean is different from 200” or “The mean is not 200.” These two hypotheses are written:

$$H_0: \mu = 200$$

$$H_1: \mu \neq 200$$

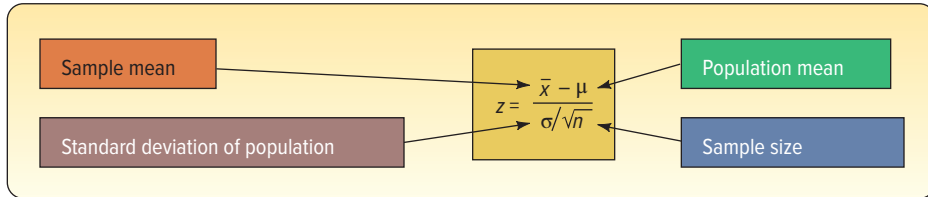
This is a *two-tailed test* because the alternate hypothesis does not state a direction. In other words, it does not state whether the mean production is greater than 200 or less than 200. The vice president wants only to find out whether the production rate is different from 200.

Before moving to Step 2, we wish to emphasize two points.

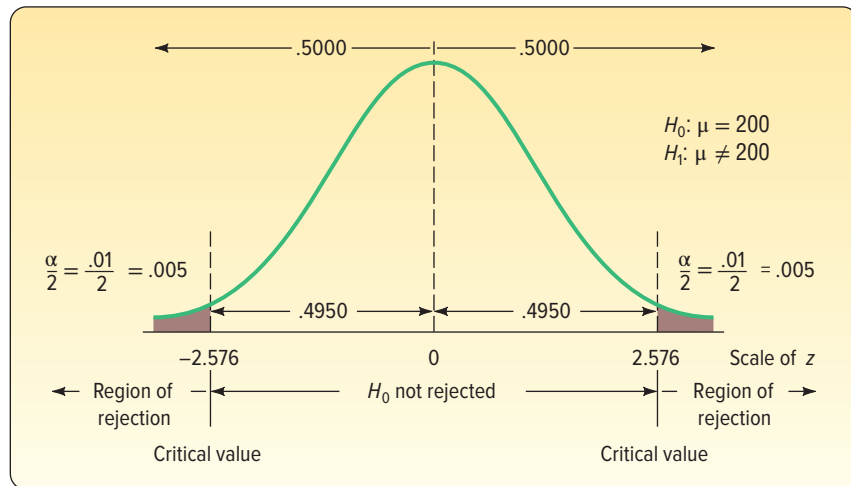
- The null hypothesis has the equal sign. Why? Because the value we are testing is always in the null hypothesis. Logically, the alternate hypothesis never contains the equal sign.
- Both the null hypothesis and the alternate hypothesis contain Greek letters—in this case  $\mu$ , which is the symbol for the population mean. Tests of hypothesis **always** refer to population parameters, never to sample statistics. To put it another way, you will never see the symbol  $\bar{x}$  as part of the null hypothesis or the alternate hypothesis.

**Step 2: Select the level of significance.** In the example description, the significance level selected is .01. This is  $\alpha$ , the probability of committing a Type I error, and it is the probability of rejecting a true null hypothesis.

**Step 3: Select the test statistic.** The test statistic is  $z$  when the population standard deviation is known. Transforming the production data to standard units ( $z$  values) permits their use not only in this problem but also in other hypothesis-testing problems. Formula (10–1) for  $z$  is repeated next with the various letters identified.



**Step 4: Formulate the decision rule.** We formulate the decision rule by first determining the critical values of  $z$ . Because this is a two-tailed test, half of .01, or .005, is placed in each tail. The area where  $H_0$  is not rejected, located between the two tails, is therefore .99. Using the Student’s  $t$  Distribution table in Appendix B.5, move to the top margin called “Level of Significance for Two-Tailed Test  $\alpha$ ,” select the column with  $\alpha = .01$ , and move to the last row, which is labeled  $\infty$ , or infinite degrees of freedom. The  $z$  value in this cell is 2.576. All the facets of this problem are shown in the diagram in Chart 10–4.



**CHART 10–4** Decision Rule for the .01 Significance Level

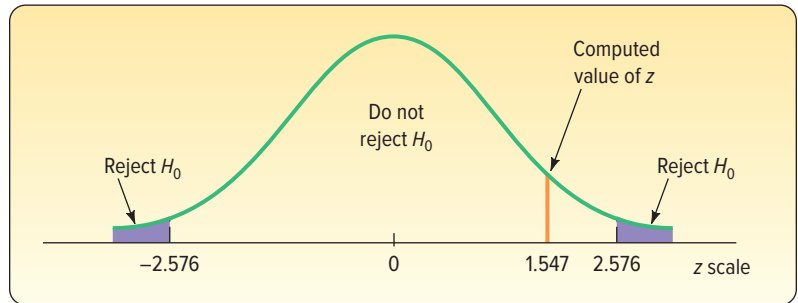
The decision rule is: If the computed value of  $z$  is not between  $-2.576$  and  $2.576$ , reject the null hypothesis. If  $z$  falls between  $-2.576$  and  $2.576$ , do not reject the null hypothesis.

**Step 5: Make a decision.** Take a sample from the population (weekly production), compute a test statistic, apply the decision rule, and arrive at a decision to reject  $H_0$  or not to reject  $H_0$ . The mean number of desks produced last year (50 weeks because the plant was shut down 2 weeks for vacation) is 203.5. The standard deviation of the population is 16 desks per week. Computing the  $z$  value from formula (10–1):

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{203.5 - 200}{16/\sqrt{50}} = 1.547$$

Because 1.547 is between  $-2.576$  and  $2.576$ , we decide not to reject  $H_0$ .

**Step 6: Interpret the result.** We did not reject the null hypothesis, so we have failed to show that the population mean has changed from 200 per week. To put it another way, the difference between the population mean of 200 per week and the sample mean of 203.5 could simply be due to chance. What should we tell the vice president? The sample information fails to indicate that the new production methods resulted in a change in the 200-desks-per-week production rate.



Did we prove that the assembly rate is still 200 per week? Not really. *We failed to disprove the null hypothesis.* Failing to disprove the hypothesis that the population mean is 200 is not the same thing as proving it to be true. For example, in the U.S. judicial system, a person is presumed innocent until proven guilty. The trial starts with a null hypothesis that the individual is innocent. If the individual is acquitted, the trial did not provide enough evidence to reject the presumption of innocence and conclude that the individual was not innocent, or guilty, as charged. That is what we do in statistical hypothesis testing when we do not reject the null hypothesis. The correct interpretation is that, based on the evidence or sample information, we have failed to disprove the null hypothesis.

We selected the significance level, .01 in this case, before setting up the decision rule and sampling the population. This is the appropriate strategy. The significance level should be set by the investigator, but it should be determined *before* gathering the sample evidence and not changed based on the sample evidence.

How does the hypothesis testing procedure just described compare with that of confidence intervals discussed in the previous chapter? When we conducted the test of hypothesis regarding the production of desks, we changed the units from desks per

week to a  $z$  value. Then we compared the computed value of the test statistic (1.547) to that of the critical values ( $-2.576$  and  $2.576$ ). Because the computed value of the test statistic was in the region where the null hypothesis was not rejected, we concluded that the population mean could be 200. To use the confidence interval approach, on the other hand, we would develop a confidence interval, based on formula (9–1). See page 247. The interval would be from 197.671 to 209.329, found by  $203.5 \pm 2.576(16/\sqrt{50})$ . Note that the proposed population value, 200, is within this interval. Hence, we would conclude that the population mean could reasonably be 200.

In general,  $H_0$  is rejected if the confidence interval does not include the hypothesized value. If the confidence interval includes the hypothesized value, then  $H_0$  is not rejected. So the “do not reject region” for a test of hypothesis is equivalent to the proposed population value occurring in the confidence interval.

## SELF-REVIEW 10–1



Heinz, a manufacturer of ketchup, uses a particular machine to dispense 16 ounces of its ketchup into containers. From many years of experience with that particular dispensing machine, Heinz knows the amount of product in each container follows a normal distribution with a mean of 16 ounces and a standard deviation of 0.15 ounce. A sample of 50 containers filled last hour revealed the mean amount per container was 16.017 ounces. Does this evidence suggest that the mean amount dispensed is different from 16 ounces? Use the .05 significance level.

- State the null hypothesis and the alternate hypothesis.
- What is the probability of a Type I error?
- Give the formula for the test statistic.
- State the decision rule.
- Determine the value of the test statistic.
- What is your decision regarding the null hypothesis?
- Interpret, in a single sentence, the result of the statistical test.



©Pamela Carley

## A One-Tailed Test

In the previous example/solution, we emphasized that we were concerned only with reporting to the vice president whether there had been a change in the mean number of desks assembled at the Fredonia plant. We were not concerned with whether the change was an increase or a decrease in the production.

To illustrate a one-tailed test, let's change the problem. Suppose the vice president wants to know whether there has been an *increase* in the number of units assembled. Can we conclude, because of the improved production methods, that the mean number of desks assembled in the last 50 weeks was more than 200? Look at the difference in the way the problem is formulated. In the first case, we wanted to know whether there was a *difference* in the mean number assembled, but now we want to know whether there has been an *increase*. Because we are investigating different questions, we will set our hypotheses differently. The biggest difference occurs in the alternate hypothesis. Before, we stated the alternate hypothesis as “different from”; now we want to state it as “greater than.” In symbols:

**A two-tailed test:**

$$H_0: \mu = 200$$

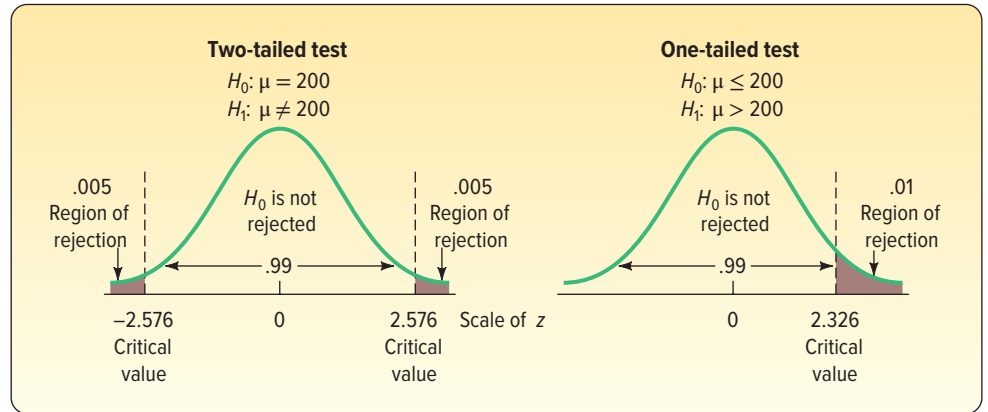
$$H_1: \mu \neq 200$$

**A one-tailed test:**

$$H_0: \mu \leq 200$$

$$H_1: \mu > 200$$

The critical values for a one-tailed test are different from a two-tailed test at the same significance level. In the previous example/solution, we split the significance level in half and put half in the lower tail and half in the upper tail. In a one-tailed test, we put all the rejection region in one tail. See Chart 10–5.



**CHART 10–5** Rejection Regions for Two-Tailed and One-Tailed Tests,  $\alpha = .01$

For the one-tailed test, the critical value of  $z$  is 2.326. Using the Student's  $t$  Distribution table in Appendix B.5, move to the top heading called “Level of Significance for One-Tailed Test,  $\alpha$ ,” select the column with  $\alpha = .01$ , and move to the last row, which is labeled  $\infty$ , or infinite degrees of freedom. The  $z$  value in this cell is 2.326.

#### LO10-5

Compute and interpret a  $p$ -value.

#### STATISTICS IN ACTION

There is a difference between *statistically significant* and *practically significant*. To explain, suppose we develop a new diet pill and test it on 100,000 people. We conclude that the typical person taking the pill for 2 years lost 1 pound. Do you think many people would be interested in taking the pill to lose 1 pound? The results of using the new pill were statistically significant but not practically significant.

## $p$ -VALUE IN HYPOTHESIS TESTING

In testing a hypothesis, we compare the test statistic to a critical value. A decision is made to either reject or not reject the null hypothesis. So, for example, if the critical value is 1.96 and the computed value of the test statistic is 2.19, the decision is to reject the null hypothesis.

In recent years, spurred by the availability of computer software, additional information is often reported on the strength of the rejection. That is, how confident are we in rejecting the null hypothesis? This approach reports the probability (assuming that the null hypothesis is true) of getting a value of the test statistic at least as extreme as the value actually obtained. This process compares the probability, called the  **$p$ -value**, with the significance level. If the  $p$ -value is smaller than the significance level,  $H_0$  is rejected. If it is larger than the significance level,  $H_0$  is not rejected.

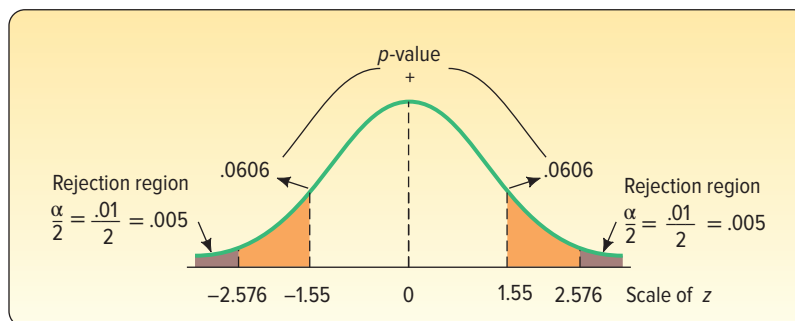
**$p$ -VALUE** The probability of observing a sample value as extreme as, or more extreme than, the value observed, given that the null hypothesis is true.

Determining the  $p$ -value not only results in a decision regarding  $H_0$ , but it gives us additional insight into the strength of the decision. A very small  $p$ -value, such as .0001, indicates that there is little likelihood the  $H_0$  is true. On the other hand, a  $p$ -value of .2033 means that  $H_0$  is not rejected, and there is little likelihood that it is false.

How do we find the  $p$ -value? To calculate  $p$ -values, we will need to use the  $z$  table (Appendix B.3) and, to use this table, we will round  $z$  test statistics to two decimals. To illustrate how to compute a  $p$ -value, we will use the example where we tested the null hypothesis that the mean number of desks produced per week at Fredonia was 200.



We did not reject the null hypothesis because the computed  $z$  test statistic of 1.547 fell in the region between  $-2.576$  and  $2.576$ . We agreed not to reject the null hypothesis if the  $z$  test statistic fell in this region. Rounding 1.547 to 1.55 and using the  $z$  table, the probability of finding a  $z$  value of 1.55 or more is .0606, found by  $.5000 - .4394$ . To put it another way, the probability of obtaining an  $\bar{x}$  greater than 203.5 if  $\mu = 200$  is .0606. To compute the  $p$ -value, we need to be concerned with the region less than  $-1.55$  as well as the values greater than 1.55 (because the rejection region is in both tails). The two-tailed  $p$ -value is .1212, found by  $2(.0606)$ . The  $p$ -value of .1212 is greater than the significance level of .01 decided upon initially, so  $H_0$  is not rejected. The details are shown in the following graph. Notice for the two-tailed hypothesis test, the  $p$ -value is represented by areas in both tails of the distribution. Then the  $p$ -value can easily be compared with the significance level. The same decision rule is used as in the one-sided test.



A  $p$ -value is a way to express the likelihood that  $H_0$  is false. But how do we interpret a  $p$ -value? We have already said that if the  $p$ -value is less than the significance level, then we reject  $H_0$ ; if it is greater than the significance level, then we do not reject  $H_0$ . Also, if the  $p$ -value is very large, then it is likely that  $H_0$  is true. If the  $p$ -value is small, then it is likely that  $H_0$  is not true. The following box will help to interpret  $p$ -values.

#### INTERPRETING THE WEIGHT OF EVIDENCE AGAINST $H_0$

If the  $p$ -value is less than

- .10, we have *some* evidence that  $H_0$  is not true.
- .05, we have *strong* evidence that  $H_0$  is not true.
- .01, we have *very strong* evidence that  $H_0$  is not true.
- .001, we have *extremely strong* evidence that  $H_0$  is not true.

## SELF-REVIEW 10-2



Refer to Self-Review 10-1.

- Suppose the next to the last sentence is changed to read: Does this evidence suggest that the mean amount dispensed is *more than* 16 ounces? State the null hypothesis and the alternate hypothesis under these conditions.
- What is the decision rule under the new conditions stated in part (a)?
- A second sample of 50 filled containers revealed the mean to be 16.040 ounces. What is the value of the test statistic for this sample?
- What is your decision regarding the null hypothesis?
- Interpret, in a single sentence, the result of the statistical test.
- What is the  $p$ -value? What is your decision regarding the null hypothesis based on the  $p$ -value? Is this the same conclusion reached in part (d)?

## EXERCISES

For Exercises 1–4, answer the questions: (a) Is this a one- or two-tailed test? (b) What is the decision rule? (c) What is the value of the test statistic? (d) What is your decision regarding  $H_0$ ? (e) What is the  $p$ -value? Interpret it.

1. A sample of 36 observations is selected from a normal population. The sample mean is 49, and the population standard deviation is 5. Conduct the following test of hypothesis using the .05 significance level.

$$H_0: \mu = 50$$

$$H_1: \mu \neq 50$$

2. A sample of 36 observations is selected from a normal population. The sample mean is 12, and the population standard deviation is 3. Conduct the following test of hypothesis using the .01 significance level.

$$H_0: \mu \leq 10$$

$$H_1: \mu > 10$$

3. A sample of 36 observations is selected from a normal population. The sample mean is 21, and the population standard deviation is 5. Conduct the following test of hypothesis using the .05 significance level.

$$H_0: \mu \leq 20$$

$$H_1: \mu > 20$$

4. A sample of 64 observations is selected from a normal population. The sample mean is 215, and the population standard deviation is 15. Conduct the following test of hypothesis using the .025 significance level.

$$H_0: \mu \geq 220$$

$$H_1: \mu < 220$$

For Exercises 5–8: (a) State the null hypothesis and the alternate hypothesis. (b) State the decision rule. (c) Compute the value of the test statistic. (d) What is your decision regarding  $H_0$ ? (e) What is the  $p$ -value? Interpret it.

5. The manufacturer of the X-15 steel-belted radial truck tire claims that the mean mileage the tire can be driven before the tread wears out is 60,000 miles. Assume the mileage wear follows the normal distribution and the standard deviation of the distribution is 5,000 miles. Crosset Truck Company bought 48 tires and found that the mean mileage for its trucks is 59,500 miles. Is Crosset's experience different from that claimed by the manufacturer at the .05 significance level?
6. The waiting time for customers at MacBurger Restaurants follows a normal distribution with a population standard deviation of 1 minute. At the Warren Road MacBurger, the quality assurance department sampled 50 customers and found that the mean waiting time was 2.75 minutes. At the .05 significance level, can we conclude that the mean waiting time is less than 3 minutes?
7. A recent national survey found that high school students watched an average (mean) of 6.8 movies per month with a population standard deviation of 1.8. The distribution of number of movies watched per month follows the normal distribution. A random sample of 36 college students revealed that the mean number of movies watched last month was 6.2. At the .05 significance level, can we conclude that college students watch fewer movies a month than high school students?
8. At the time she was hired as a server at the Grumney Family Restaurant, Beth Brigden was told, "You can average \$80 a day in tips." Assume the population of daily tips is normally distributed with a standard deviation of \$9.95. Over the first 35 days she was employed at the restaurant, the mean daily amount of her tips was \$84.85. At the .01 significance level, can Ms. Brigden conclude that her daily tips average more than \$80?

**LO10-6**

Use a  $t$  statistic to test a hypothesis.

## HYPOTHESIS TESTING FOR A POPULATION MEAN: POPULATION STANDARD DEVIATION UNKNOWN

In the preceding example, we knew  $\sigma$ , the population standard deviation, and that the population followed the normal distribution. In most cases, however, the population standard deviation is unknown. Thus,  $\sigma$  must be based on prior studies or estimated by the sample standard deviation,  $s$ . The population standard deviation in the following example is not known, so the sample standard deviation is used to estimate  $\sigma$ .

To find the value of the test statistic, we use the  $t$  distribution and revise formula (10–1) as follows:

**TESTING A MEAN,  $\sigma$  UNKNOWN**

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

**(10–2)**

with  $n - 1$  degrees of freedom, where:

$\bar{x}$  is the sample mean.

$\mu$  is the hypothesized population mean.

$s$  is the sample standard deviation.

$n$  is the number of observations in the sample.

We encountered this same situation when constructing confidence intervals in the previous chapter. See pages 252–259 in Chapter 9. We summarized this problem in Chart 9–3 on page 254. Under these conditions, the correct statistical procedure is to replace the standard normal distribution with the  $t$  distribution. To review, the major characteristics of the  $t$  distribution are:

- It is a continuous distribution.
- It is bell-shaped and symmetrical.
- There is a family of  $t$  distributions. Each time the degrees of freedom change, a new distribution is created.
- As the number of degrees of freedom increases, the shape of the  $t$  distribution approaches that of the standard normal distribution.
- The  $t$  distribution is flatter, or more spread out, than the standard normal distribution.

The following example/solution shows the details.

### ▶ EXAMPLE

The McFarland Insurance Company Claims Department reports the mean cost to process a claim is \$60. An industry comparison showed this amount to be larger than most other insurance companies, so the company instituted cost-cutting measures. To evaluate the effect of the cost-cutting measures, the supervisor of the Claims Department selected a random sample of 26 claims processed last month and recorded the cost to process each claim. The sample information is reported below.

|      |      |      |      |      |      |
|------|------|------|------|------|------|
| \$45 | \$49 | \$62 | \$40 | \$43 | \$61 |
| 48   | 53   | 67   | 63   | 78   | 64   |
| 48   | 54   | 51   | 56   | 63   | 69   |
| 58   | 51   | 58   | 59   | 56   | 57   |
| 38   | 76   |      |      |      |      |

At the .01 significance level, is it reasonable to conclude that the mean cost to process a claim is now less than \$60?

### SOLUTION

We will use the six-step hypothesis testing procedure.

**Step 1: State the null hypothesis and the alternate hypothesis.** The null hypothesis is that the population mean is at least \$60. The alternate hypothesis is that the population mean is less than \$60. We can express the null and alternate hypotheses as follows:

$$H_0: \mu \geq \$60$$

$$H_1: \mu < \$60$$

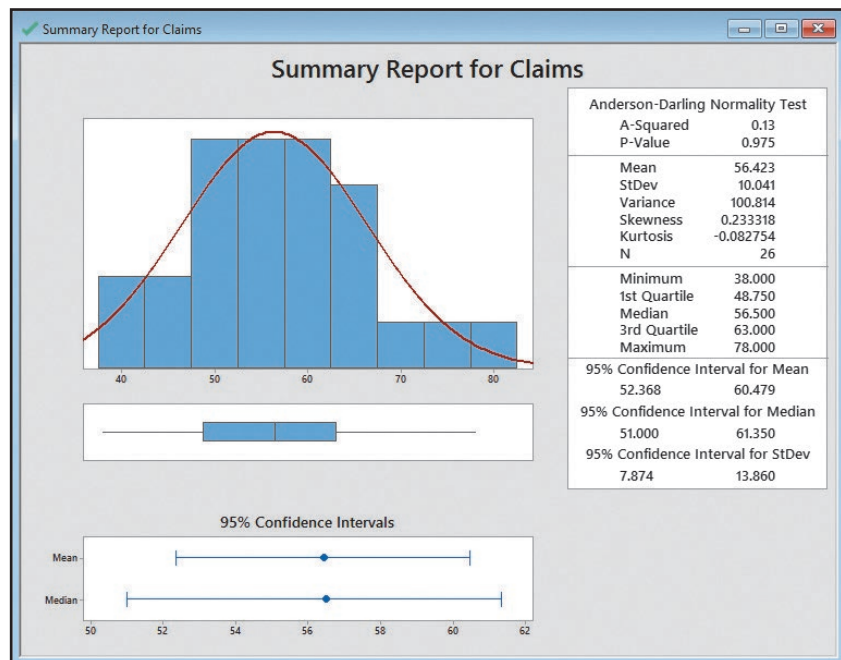
The test is *one-tailed* because we want to determine whether there has been a *reduction* in the cost. The inequality in the alternate hypothesis points to the region of rejection in the left tail of the distribution.

**Step 2: Select the level of significance.** We decided on the .01 significance level.

**Step 3: Select the test statistic.** The test statistic in this situation is the *t* distribution. Why? First, it is reasonable to conclude that the distribution of the cost per claim follows the normal distribution. We can confirm this from the histogram in the center of the following Minitab output. Observe the normal distribution superimposed on the frequency distribution.

We do not know the standard deviation of the population. So we substitute the sample standard deviation. The value of the test statistic is computed by formula (10–2):

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$



Source: Minitab

TABLE 10–1 A Portion of the  $t$  Distribution Table

| Confidence Intervals |   |       |       |       |       |        |
|----------------------|---|-------|-------|-------|-------|--------|
|                      | 80%   | 90%   | 95%   | 98%   | 99%   | 99.9%  |
| df                   | Level of Significance for One-Tailed Test, $\alpha$ |       |       |       |       |        |
|                      | 0.10  | 0.05  | 0.025 | 0.01  | 0.005 | 0.0005 |
|                      | Level of Significance for Two-Tailed Test, $\alpha$ |       |       |       |       |        |
|                      | 0.20  | 0.10  | 0.05  | 0.02  | 0.01  | 0.001  |
| ∴                    | ∴   | ∴     | ∴     | ∴     | ∴     | ∴      |
| 21                   | 1.323   | 1.721 | 2.080 | 2.518 | 2.831 | 3.819  |
| 22                   | 1.321   | 1.717 | 2.074 | 2.508 | 2.819 | 3.792  |
| 23                   | 1.319   | 1.714 | 2.069 | 2.500 | 2.807 | 3.768  |
| 24                   | 1.318   | 1.711 | 2.064 | 2.492 | 2.797 | 3.745  |
| 25                   | 1.316   | 1.708 | 2.060 | 2.485 | 2.787 | 3.725  |
| 26                   | 1.315   | 1.706 | 2.056 | 2.479 | 2.779 | 3.707  |
| 27                   | 1.314   | 1.703 | 2.052 | 2.473 | 2.771 | 3.690  |
| 28                   | 1.313   | 1.701 | 2.048 | 2.467 | 2.763 | 3.674  |
| 29                   | 1.311   | 1.699 | 2.045 | 2.462 | 2.756 | 3.659  |
| 30                   | 1.310   | 1.697 | 2.042 | 2.457 | 2.750 | 3.646  |

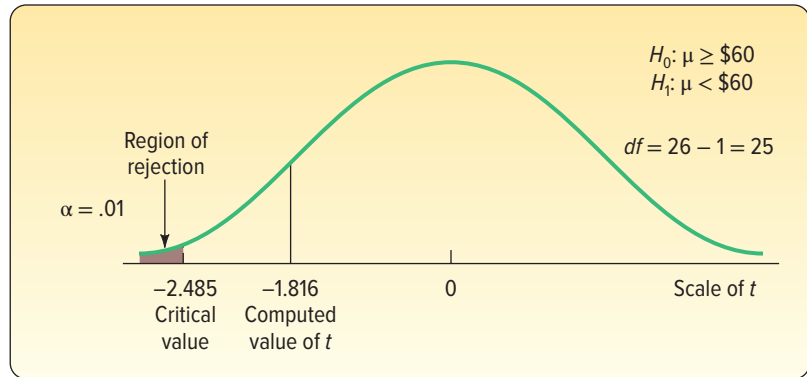
**Step 4: Formulate the decision rule.** The critical values of  $t$  are given in Appendix B.5, a portion of which is shown in Table 10–1. The far left column of the table is labeled “ $df$ ” for degrees of freedom. The number of degrees of freedom is the total number of observations in the sample minus the number of populations sampled, written  $n - 1$ . In this case, the number of observations in the sample is 26, and we sampled 1 population, so there are  $26 - 1 = 25$  degrees of freedom. To find the critical value, first locate the row with the appropriate degrees of freedom. This row is shaded in Table 10–1. Next, determine whether the test is one-tailed or two-tailed. In this case, we have a one-tailed test, so find the portion of the table that is labeled “one-tailed.” Locate the column with the selected significance level. In this example, the significance level is .01. Move down the column labeled “0.01” until it intersects the row with 25 degrees of freedom. The value is 2.485. Because this is a one-sided test and the rejection region is in the left tail, the critical value is negative. The decision rule is to reject  $H_0$  if the value of  $t$  is less than  $-2.485$ .

**Step 5: Make a decision.** From the Minitab output, the mean cost per claim for the sample of 26 observations is \$56.423. The standard deviation of this sample is \$10.041. We insert these values in formula (10–2) and compute the value of  $t$ :

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\$56.423 - \$60}{\$10.041/\sqrt{26}} = -1.816$$

Because  $-1.816$  lies in the region to the right of the critical value of  $-2.485$  (see Chart 10–6), the null hypothesis is not rejected at the .01 significance level.

**Step 6: Interpret the result.** We have not disproved the null hypothesis. The sample of claims could have been selected from a population with a



**CHART 10-6** Rejection Region,  $t$  Distribution, .01 Significance Level

mean cost of \$60 per claim. To put it another way, the difference of \$3.577 (\$56.423 – \$60.00) between the sample mean and the population mean could be due to sampling error. The test results do not allow the Claims Department manager to conclude that the cost-cutting measures have been effective.

In the previous example, the mean and the standard deviation were computed using Minitab. The following example/solution shows the details when the sample mean and sample standard deviation are calculated from sample data.

### EXAMPLE

The Myrtle Beach International Airport provides a cell phone parking lot where people can wait for a message to pick up arriving passengers. To decide if the cell phone lot has enough parking places, the manager of airport parking needs to know if the mean time in the lot is more than 15 minutes. A sample of 12 recent customers showed they were in the lot the following lengths of time, in minutes.

30   24   28   22   14   2   39   23   23   28   12   31

At the .05 significance level, is it reasonable to conclude that the mean time in the lot is more than 15 minutes?

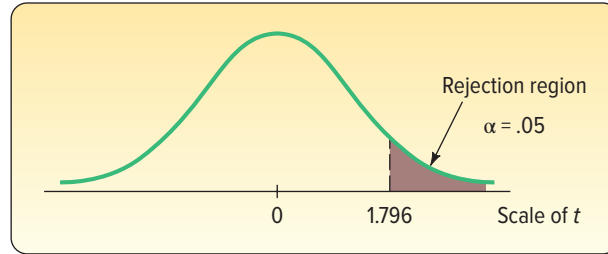
### SOLUTION

We begin by stating the null hypothesis and the alternate hypothesis. In this case, the question is whether the population mean could be more than 15 minutes. So this is a one-tailed test. We state the two hypotheses as follows:

$$H_0: \mu \leq 15$$

$$H_1: \mu > 15$$

There are 11 degrees of freedom, found by  $n - 1 = 12 - 1 = 11$ . The critical  $t$  value is 1.796, found by referring to Appendix B.5 for a one-tailed test, using  $\alpha = .05$  with 11 degrees of freedom. The decision rule is: Reject the null hypothesis if the computed  $t$  is greater than 1.796. This information is summarized in Chart 10-7.



**CHART 10–7** Rejection Region, One-Tailed Test, Student’s *t* Distribution,  $\alpha = .05$

**TABLE 10–2** Calculations of Sample Mean and Standard Deviation Parking Times

| Customer | <i>x</i> , Minutes | $(x - \bar{x})^2$ |
|----------|--------------------|-------------------|
| Chmura   | 30                 | 49                |
| Will     | 24                 | 1                 |
| Crompton | 28                 | 25                |
| Craver   | 22                 | 1                 |
| Cao      | 14                 | 81                |
| Nowlin   | 2                  | 441               |
| Esposito | 39                 | 256               |
| Colvard  | 23                 | 0                 |
| Hoeffe   | 23                 | 0                 |
| Lawler   | 28                 | 25                |
| Trask    | 12                 | 121               |
| Grullon  | 31                 | 64                |
| Total    | 276                | 1064              |

$$\bar{x} = \frac{\sum x}{n} = \frac{276}{12} = 23$$

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{1064}{12 - 1}} = 9.835$$

We calculate the sample mean using formula (3–2) and the sample standard deviation using formula (3–8). The sample mean is 23 minutes, and the sample standard deviation is 9.835 minutes. The details of the calculations are shown in Table 10–2.

Now we are ready to compute the value of *t*, using formula (10–2).

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{23 - 15}{9.835/\sqrt{12}} = 2.818$$

The null hypothesis that the population mean is less than or equal to 15 minutes is rejected because the computed *t* value of 2.818 lies in the area to the right of 1.796. We conclude that the time customers spend in the lot is more than 15 minutes. This result indicates that the airport may need to add more parking places.

### SELF-REVIEW 10–3



The mean life of a battery used in a digital clock is 305 days. The lives of the batteries follow the normal distribution. The battery was recently modified to last longer. A sample of 20 of the modified batteries had a mean life of 311 days with a standard deviation of 12 days. Did the modification increase the mean life of the battery?

- State the null hypothesis and the alternate hypothesis.
- Show the decision rule graphically. Use the .05 significance level.
- Compute the value of *t*. What is your decision regarding the null hypothesis? Briefly summarize your results.

## EXERCISES

9. Given the following hypotheses:

$$H_0: \mu \leq 10$$

$$H_1: \mu > 10$$

A random sample of 10 observations is selected from a normal population. The sample mean was 12 and the sample standard deviation was 3. Using the .05 significance level:

- State the decision rule.
  - Compute the value of the test statistic.
  - What is your decision regarding the null hypothesis?
10. Given the following hypotheses:

$$H_0: \mu = 400$$

$$H_1: \mu \neq 400$$

A random sample of 12 observations is selected from a normal population. The sample mean was 407 and the sample standard deviation was 6. Using the .01 significance level:

- State the decision rule.
  - Compute the value of the test statistic.
  - What is your decision regarding the null hypothesis?
11. The Rocky Mountain district sales manager of Rath Publishing Inc., a college textbook publishing company, claims that the sales representatives make an average of 40 sales calls per week on professors. Several reps say that this estimate is too low. To investigate, a random sample of 28 sales representatives reveals that the mean number of calls made last week was 42. The standard deviation of the sample is 2.1 calls. Using the .05 significance level, can we conclude that the mean number of calls per salesperson per week is more than 40?
12. The management of White Industries is considering a new method of assembling its golf cart. The present method requires a mean time of 42.3 minutes to assemble a cart. The mean assembly time for a random sample of 24 carts, using the new method, was 40.6 minutes, and the standard deviation of the sample was 2.7 minutes. Using the .10 level of significance, can we conclude that the assembly time using the new method is faster?
13. The mean income per person in the United States is \$50,000, and the distribution of incomes follows a normal distribution. A random sample of 10 residents of Wilmington, Delaware, had a mean of \$60,000 with a standard deviation of \$10,000. At the .05 level of significance, is that enough evidence to conclude that residents of Wilmington, Delaware, have more income than the national average?
14. **FILE** Most air travelers now use e-tickets. Electronic ticketing allows passengers to not worry about a paper ticket, and it costs the airline companies less than paper ticketing. However, recently the airlines have received complaints from passengers regarding their e-tickets, particularly when connecting flights and a change of airlines were involved. To investigate the problem, an independent watchdog agency contacted a random sample of 20 airports and collected information on the number of complaints the airport had with e-tickets for the month of March. The information is reported below.

|    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|
| 14 | 14 | 16 | 12 | 12 | 14 | 13 | 16 | 15 | 14 |
| 12 | 15 | 15 | 14 | 13 | 13 | 12 | 13 | 10 | 13 |

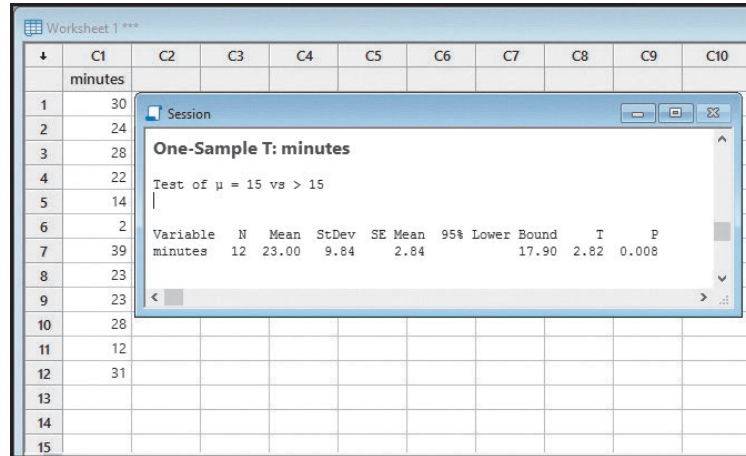
At the .05 significance level, can the watchdog agency conclude the mean number of complaints per airport is less than 15 per month?

- What assumption is necessary before conducting a test of hypothesis?
- Plot the number of complaints per airport in a frequency distribution or a dot plot. Is it reasonable to conclude that the population follows a normal distribution?
- Conduct a test of hypothesis and interpret the results.



## A Statistical Software Solution

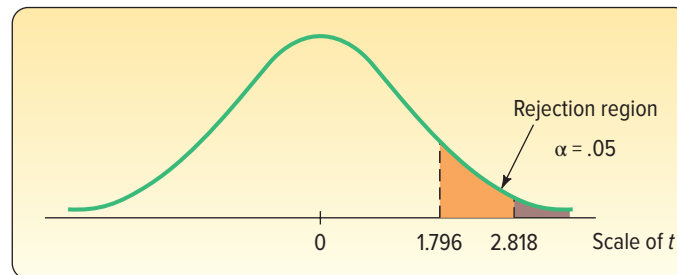
The Minitab statistical software system, used in earlier chapters and the previous section, provides an efficient method for conducting a one-sample test of hypothesis for a population mean. The steps to generate the following output are shown in Appendix C.



Source: Minitab

An additional feature of most statistical software packages is to report the  $p$ -value, which gives additional information on the null hypothesis. The  $p$ -value is the probability of a  $t$  value as extreme or more extreme than the computed  $t$  value, given that the null hypothesis is true. Using the Minitab analysis from the previous cell phone parking lot example, the  $p$ -value of .008 is the likelihood of a  $t$  value of 2.82 or larger, given a population mean of 15. Thus, comparing the  $p$ -value to the significance level tells us whether the null hypothesis was close to being rejected, barely rejected, and so on.

To explain further, refer to the diagram below. The  $p$ -value of .008 is the brown shaded area and the significance level is the total amber and brown shaded area. Because the  $p$ -value of .008 is less than the significance level of .05, the null hypothesis is rejected. Had the  $p$ -value been larger—say, .06, .19, or .57—than the significance level, the null hypothesis would not be rejected.



In the preceding example, the alternate hypothesis was one-sided, and the upper (right) tail of the  $t$  distribution contained the rejection region. The  $p$ -value is the area to the right of 2.818 for a  $t$  distribution with 11 degrees of freedom.

What if we were conducting a two-sided test, so that the rejection region is in both the upper and the lower tails? That is, in the cell phone parking lot example, if  $H_1$  were stated as  $\mu \neq 15$ , we would have reported the  $p$ -value as the area to the right of 2.818 plus the value to the left of  $-2.818$ . Both of these values are .008, so the  $p$ -value is  $.008 + .008 = .016$ .

How can we estimate a  $p$ -value without a computer? To illustrate, recall that, in the example/solution regarding the length of time at the cell phone parking lot, we

rejected the null hypothesis that  $\mu \leq 15$  and accepted the alternate hypothesis that  $\mu > 15$ . The significance level was .05, so logically the  $p$ -value is less than .05. To estimate the  $p$ -value more accurately, go to Appendix B.5 and find the row with 11 degrees of freedom. The computed  $t$  value of 2.818 is between 2.718 and 3.106. (A portion of Appendix B.5 is reproduced as Table 10–3.) The one-tailed significance level corresponding to 2.718 is .01, and for 3.106 it is .005. Therefore, the  $p$ -value is between .005 and .01. The usual practice is to report that the  $p$ -value is less than the larger of the two significance levels. So we would report “the  $p$ -value is less than .01.”

**TABLE 10–3** A Portion of Student’s  $t$  Distribution

| Confidence Intervals |   |       |       |       |       |        |
|----------------------|---|-------|-------|-------|-------|--------|
|                      | 80%   | 90%   | 95%   | 98%   | 99%   | 99.9%  |
| df                   | Level of Significance for One-Tailed Test, $\alpha$ |       |       |       |       |        |
|                      | 0.10  | 0.05  | .0025 | 0.01  | 0.005 | 0.0005 |
|                      | Level of Significance for Two-Tailed Test, $\alpha$ |       |       |       |       |        |
|                      | 0.20  | 0.10  | 0.05  | 0.02  | 0.01  | 0.001  |
| ∴                    | ∴   | ∴     | ∴     | ∴     | ∴     | ∴      |
| 9                    | 1.383   | 1.833 | 2.262 | 2.821 | 3.250 | 4.781  |
| 10                   | 1.372   | 1.812 | 2.228 | 2.764 | 3.169 | 4.587  |
| 11                   | 1.363   | 1.796 | 2.201 | 2.718 | 3.106 | 4.437  |
| 12                   | 1.356   | 1.782 | 2.179 | 2.681 | 3.055 | 4.318  |
| 13                   | 1.350   | 1.771 | 2.160 | 2.650 | 3.012 | 4.221  |
| 14                   | 1.345   | 1.761 | 2.145 | 2.624 | 2.977 | 4.140  |
| 15                   | 1.341   | 1.753 | 2.131 | 2.602 | 2.947 | 4.073  |

**SELF-REVIEW 10–4**



A machine is set to fill a small bottle with 9.0 grams of medicine. A sample of eight bottles revealed the following amounts (grams) in each bottle.

- 9.2   8.7   8.9   8.6   8.8   8.5   8.7   9.0

- At the .01 significance level, can we conclude that the mean weight is less than 9.0 grams?
- State the null hypothesis and the alternate hypothesis.
  - How many degrees of freedom are there?
  - Give the decision rule.
  - Compute the value of  $t$ . What is your decision regarding the null hypothesis?
  - Estimate the  $p$ -value.

**EXERCISES**

15. Given the following hypotheses:

$$H_0: \mu \geq 20$$

$$H_1: \mu > 20$$

A random sample of five resulted in the following values: 18, 15, 12, 19, and 21. Assume a normal population. Using the .01 significance level, can we conclude the population mean is less than 20?

- State the decision rule.
- Compute the value of the test statistic.
- What is your decision regarding the null hypothesis?
- Estimate the  $p$ -value.

16. Given the following hypotheses:

$$H_0: \mu = 100$$

$$H_1: \mu \neq 100$$

A random sample of six resulted in the following values: 118, 105, 112, 119, 105, and 111. Assume a normal population. Using the .05 significance level, can we conclude the mean is different from 100?

- State the decision rule.
  - Compute the value of the test statistic.
  - What is your decision regarding the null hypothesis?
  - Estimate the  $p$ -value.
17. **FILE** The amount of water consumed each day by a healthy adult follows a normal distribution with a mean of 1.4 liters. A health campaign promotes the consumption of at least 2.0 liters per day. A sample of 10 adults after the campaign shows the following consumption in liters:

1.5    1.6    1.5    1.4    1.9    1.4    1.3    1.9    1.8    1.7

At the .01 significance level, can we conclude that water consumption has increased? Calculate and interpret the  $p$ -value.

18. **FILE** The liquid chlorine added to swimming pools to combat algae has a relatively short shelf life before it loses its effectiveness. Records indicate that the mean shelf life of a 5-gallon jug of chlorine is 2,160 hours (90 days). As an experiment, Holdlonger was added to the chlorine to find whether it would increase the shelf life. A sample of nine jugs of chlorine had these shelf lives (in hours):

2,159    2,170    2,180    2,179    2,160    2,167    2,171    2,181    2,185

At the .025 level, has Holdlonger increased the shelf life of the chlorine? Estimate the  $p$ -value.

19. **FILE** A Washington, D.C., “think tank” announces the typical teenager sent 67 text messages per day in 2017. To update that estimate, you phone a sample of 12 teenagers and ask them how many text messages they sent the previous day. Their responses were:

51    175    47    49    44    54    145    203    21    59    42    100

At the .05 level, can you conclude that the mean number is greater than 67? Estimate the  $p$ -value and describe what it tells you.

20. **FILE** Hugger Polls contends that an agent conducts a mean of 53 in-depth home surveys every week. A streamlined survey form has been introduced, and Hugger wants to evaluate its effectiveness. The number of in-depth surveys conducted during a week by a random sample of 15 agents are:

53    57    50    55    58    54    60    52    59    62    60    60    51    59    56

At the .05 level of significance, can we conclude that the mean number of interviews conducted by the agents is more than 53 per week? Estimate the  $p$ -value.

## CHAPTER SUMMARY

- I. The objective of hypothesis testing is to verify the validity of a statement about a population parameter.
- II. The steps to conduct a test of hypothesis are:
  - A. State the null hypothesis ( $H_0$ ) and the alternate hypothesis ( $H_1$ ).
  - B. Select the level of significance.
    1. The level of significance is the likelihood or probability of rejecting a true null hypothesis.
    2. The most frequently used significance levels are .01, .05, and .10. As a probability, any value between 0 and 1.00 is possible, but we prefer small probabilities of making a Type I error.
  - C. Select the test statistic.
    1. A test statistic is a value calculated from sample information used to determine whether to reject the null hypothesis.
    2. Two test statistics were considered in this chapter.
      - a. The standard normal distribution (the  $z$  distribution) is used when the population follows the normal distribution and the population standard deviation is known.
      - b. The  $t$  distribution is used when the population follows the normal distribution and the population standard deviation is unknown.
  - D. State the decision rule.
    1. The decision rule indicates the condition or conditions when the null hypothesis is rejected.
    2. In a two-tailed test, the rejection region is evenly split between the upper and lower tails.
    3. In a one-tailed test, all of the rejection region is in either the upper or the lower tail.
  - E. Select a sample, compute the value of the test statistic, and make a decision regarding the null hypothesis.
  - F. Interpret the results of your decision.
- III. A  $p$ -value is the probability that the value of the test statistic is as extreme as the value computed, when the null hypothesis is true.
- IV. When testing a hypothesis about a population mean:
  - A. If the population standard deviation,  $\sigma$ , is known, the test statistic is the standard normal distribution and is determined from:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad (10-1)$$

- B. If the population standard deviation is not known,  $s$  is substituted for  $\sigma$ . The test statistic is the  $t$  distribution, and its value is determined from:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad (10-2)$$

The major characteristics of the  $t$  distribution are:

1. It is a continuous distribution.
2. It is mound-shaped and symmetrical.
3. It is flatter, or more spread out, than the standard normal distribution.
4. There is a family of  $t$  distributions, depending on the number of degrees of freedom.

## PRONUNCIATION KEY

| SYMBOL     | MEANING                       | PRONUNCIATION             |
|------------|-------------------------------|---------------------------|
| $H_0$      | Null hypothesis               | <i>H sub zero</i>         |
| $H_1$      | Alternate hypothesis          | <i>H sub one</i>          |
| $\alpha/2$ | Two-tailed significance level | <i>Alpha divided by 2</i> |

## CHAPTER EXERCISES

21. According to the local union president, the mean gross income of plumbers in the Salt Lake City area follows the normal probability distribution with a mean of \$45,000 and a standard deviation of \$3,000. A recent investigative reporter for KYAK TV found, for a sample of 120 plumbers, the mean gross income was \$45,500. At the .10 significance level, is it reasonable to conclude that the mean income is not equal to \$45,000? Determine the  $p$ -value.
22. **FILE** Rutter Nursery Company packages its pine bark mulch in 50-pound bags. From a long history, the production department reports that the distribution of the bag weights follows the normal distribution, and the standard deviation of the packaging process is 3 pounds per bag. At the end of each day, Jeff Rutter, the production manager, weighs 10 bags and computes the mean weight of the sample. Below are the weights of 10 bags from today's production.

|      |      |      |      |      |      |      |      |      |      |
|------|------|------|------|------|------|------|------|------|------|
| 45.6 | 47.7 | 47.6 | 46.3 | 46.2 | 47.4 | 49.2 | 55.8 | 47.5 | 48.5 |
|------|------|------|------|------|------|------|------|------|------|

- a. Can Mr. Rutter conclude that the mean weight of the bags is less than 50 pounds? Use the .01 significance level.
- b. In a brief report, tell why Mr. Rutter can use the  $z$  distribution as the test statistic.
- c. Compute the  $p$ -value.
23. A new weight-watching company, Weight Reducers International, advertises that those who join will lose an average of 10 pounds after the first two weeks. The standard deviation is 2.8 pounds. A random sample of 50 people who joined the weight reduction program revealed a mean loss of 9 pounds. At the .05 level of significance, can we conclude that those joining Weight Reducers will lose less than 10 pounds? Determine the  $p$ -value.
24. Dole Pineapple Inc. is concerned that the 16-ounce can of sliced pineapple is being overfilled. Assume the standard deviation of the process is .03 ounce. The quality control department took a random sample of 50 cans and found that the arithmetic mean weight was 16.05 ounces. At the 5% level of significance, can we conclude that the mean weight is greater than 16 ounces? Determine the  $p$ -value.
25. According to a recent survey, Americans get a mean of 7 hours of sleep per night. A random sample of 50 students at West Virginia University revealed the mean length of time slept last night was 6 hours and 48 minutes (6.8 hours). The standard deviation of the sample was 0.9 hour. At the 5% level of significance, is it reasonable to conclude that students at West Virginia sleep less than the typical American? Compute the  $p$ -value.
26. A statewide real estate sales agency, Farm Associates, specializes in selling farm property in the state of Nebraska. Its records indicate that the mean selling time of farm property is 90 days. Because of recent drought conditions, the agency believes that the mean selling time is now greater than 90 days. A statewide survey of 100 recently sold farms revealed a mean selling time of 94 days, with a standard deviation of 22 days. At the .10 significance level, has there been an increase in selling time?
27. According to the Census Bureau, 3.13 people reside in the typical American household. A sample of 25 households in Arizona retirement communities showed the mean number of residents per household was 2.86 residents. The standard deviation of this sample was 1.20 residents. At the .05 significance level, is it reasonable to conclude the mean number of residents in the retirement community household is less than 3.13 persons?
28. A recent article in *Vitality* magazine reported that the mean amount of leisure time per week for American men is 40.0 hours. You believe this figure is too large and decide to conduct your own test. In a random sample of 60 men, you find that the mean is 37.8 hours of leisure per week and that the standard deviation of the sample is 12.2 hours. Can you conclude that the information in the article is untrue? Use the .05 significance level. Determine the  $p$ -value and explain its meaning.

29. **FILE** A recent survey by [nerdwallet.com](http://nerdwallet.com) indicated Americans paid a mean of \$6,658 interest on credit card debt in 2017. A sample of 12 households with children revealed the following amounts. At the .05 significance level, is it reasonable to conclude that these households paid more interest?

7,077 5,744 6,753 7,381 7,625 6,636 7,164 7,348 8,060 5,848 9,275 7,052

30. **FILE** A recent article in *The Wall Street Journal* reported that the home equity loan rate is now less than 4%. A sample of eight small banks in the Midwest revealed the following home equity loan rates (in percent):

3.6 4.1 5.3 3.6 4.9 4.6 5.0 4.4

At the .01 significance level, can we conclude that the home equity loan rate for small banks is less than 4%? Estimate the  $p$ -value.

31. **FILE** A recent study revealed the typical American coffee drinker consumes an average of 3.1 cups per day. A sample of 12 senior citizens revealed they consumed the following amounts of coffee, reported in cups, yesterday.

3.1 3.3 3.5 2.6 2.6 4.3 4.4 3.8 3.1 4.1 3.1 3.2

At the .05 significance level, do these sample data suggest there is a difference between the national average and the sample mean from senior citizens?

32. **FILE** The postanesthesia care area (recovery room) at St. Luke's Hospital in Maumee, Ohio, was recently enlarged. The hope was that the change would increase the mean number of patients served per day to more than 25. A random sample of 15 days revealed the following numbers of patients.

25 27 25 26 25 28 28 27 24 26 25 29 25 27 24

At the .01 significance level, can we conclude that the mean number of patients per day is more than 25? Estimate the  $p$ -value and interpret it.

33. **FILE** [www.golfsmith.com](http://www.golfsmith.com) receives an average of 6.5 returns per day from online shoppers. For a sample of 12 days, it received the following numbers of returns.

0 4 3 4 9 4 5 9 1 6 7 10

At the .01 significance level, can we conclude the mean number of returns is less than 6.5?

34. **FILE** During recent seasons, Major League Baseball has been criticized for the length of the games. A report indicated that the average game lasts 3 hours and 30 minutes. A sample of 17 games revealed the following times to completion. (Note that the minutes have been changed to fractions of hours, so that a game that lasted 2 hours and 24 minutes is reported at 2.40 hours.)

2.98 2.40 2.70 2.25 3.23 3.17 2.93 3.18 2.80  
2.38 3.75 3.20 3.27 2.52 2.58 4.45 2.45

Can we conclude that the mean time for a game is less than 3.50 hours? Use the .05 significance level.

35. **FILE** Watch Corporation of Switzerland claims that its watches on average will neither gain nor lose time during a week. A sample of 18 watches provided the following gains (+) or losses (–) in seconds per week.

|       |       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| –0.38 | –0.20 | –0.38 | –0.32 | +0.32 | –0.23 | +0.30 | +0.25 | –0.10 |
| –0.37 | –0.61 | –0.48 | –0.47 | –0.64 | –0.04 | –0.20 | –0.68 | +0.05 |

Is it reasonable to conclude that the mean gain or loss in time for the watches is 0? Use the .05 significance level. Estimate the  $p$ -value.

36. **FILE** Listed below is the annual rate of return (reported in percent) for a sample of 12 taxable mutual funds.

|      |      |      |      |      |      |      |      |      |      |      |      |
|------|------|------|------|------|------|------|------|------|------|------|------|
| 4.63 | 4.15 | 4.76 | 4.70 | 4.65 | 4.52 | 4.70 | 5.06 | 4.42 | 4.51 | 4.24 | 4.52 |
|------|------|------|------|------|------|------|------|------|------|------|------|

Using the .05 significance level, is it reasonable to conclude that the mean rate of return is more than 4.50%?

37. **FILE** Many grocery stores and large retailers such as Kroger and Walmart have installed self-checkout systems so shoppers can scan their own items and cash out themselves. How do customers like this service and how often do they use it? Listed below is the number of customers using the service for a sample of 15 days at a Walmart location.

|     |     |     |     |     |    |     |    |     |     |
|-----|-----|-----|-----|-----|----|-----|----|-----|-----|
| 120 | 108 | 120 | 114 | 118 | 91 | 118 | 92 | 104 | 104 |
| 112 | 97  | 118 | 108 | 117 |    |     |    |     |     |

Is it reasonable to conclude that the mean number of customers using the self-checkout system is more than 100 per day? Use the .05 significance level.

38. **FILE** For a recent year, the mean fare to fly from Charlotte, North Carolina, to Chicago, Illinois, on a discount ticket was \$267. A random sample of 13 round-trip discount fares on this route last month shows:

|       |       |       |       |       |       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| \$321 | \$286 | \$290 | \$330 | \$310 | \$250 | \$270 | \$280 | \$299 | \$265 | \$291 | \$275 | \$281 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|

At the .01 significance level, can we conclude that the mean fare has increased? What is the  $p$ -value?

39. The publisher of *Celebrity Living* claims that the mean sales for personality magazines that feature people such as Megan Fox or Jennifer Lawrence are 1.5 million copies per week. A sample of 10 comparable titles shows a mean weekly sales last week of 1.3 million copies with a standard deviation of 0.9 million copies. Do these data contradict the publisher's claim? Use the .01 significance level.
40. A United Nations report shows the mean family income for Mexican migrants to the United States is \$27,000 per year. A FLOC (Farm Labor Organizing Committee) evaluation of 25 Mexican family units reveals the mean to be \$30,000 with a sample standard deviation of \$10,000. Does this information disagree with the United Nations report? Apply the .01 significance level.
41. **FILE** The number of “destination weddings” has skyrocketed in recent years. For example, many couples are opting to have their weddings in the Caribbean. A Caribbean vacation resort recently advertised in *Bride Magazine* that the cost of a Caribbean wedding was less than \$30,000. Listed below is a total cost in \$000 for a sample of eight Caribbean weddings.

|      |      |      |      |      |      |      |      |
|------|------|------|------|------|------|------|------|
| 29.7 | 29.4 | 31.7 | 29.0 | 29.1 | 30.5 | 29.1 | 29.8 |
|------|------|------|------|------|------|------|------|

At the .05 significance level, is it reasonable to conclude the mean wedding cost is less than \$30,000 as advertised?

42. The American Water Works Association reports that the per capita water use in a single-family home is 69 gallons per day. Legacy Ranch is a relatively new housing development. The builders installed more efficient water fixtures, such as low-flush toilets, and subsequently conducted a survey of the residences. Thirty-six owners responded, and the sample mean water use per day was 64 gallons with a standard deviation of 8.8 gallons per day. At the .10 level of significance, is that enough evidence to conclude that residents of Legacy Ranch use less water on average?
43. A cola-dispensing machine is set to dispense 9.00 ounces of cola per cup, with a standard deviation of 1.00 ounce. The manufacturer of the machine would like to set the control limit in such a way that, for samples of 36, 5% of the sample means will be greater than the upper control limit, and 5% of the sample means will be less than the lower control limit.
- At what value should the control limit be set?
  - If the population mean shifts to 8.6, what is the probability of detecting the change?
  - If the population mean shifts to 9.6, what is the probability of detecting the change?

## DATA ANALYTICS

(The data for these exercises are available in Connect.)

44. **FILE** The North Valley Real Estate data report information on the homes sold last year.
- Adam Marty recently joined North Valley Real Estate and was assigned 20 homes to market and show. When he was hired, North Valley assured him that the 20 homes would be fairly assigned to him. When he reviewed the selling prices of his assigned homes, he thought that the prices were much below the average of \$357,000. Adam was able to find the data of how the other agents in the firm were assigned to the homes. Use statistical inference to analyze the “fairness” of assigning homes to agents.
45. **FILE** Refer to the Baseball 2016 data, which report information on the 30 Major League Baseball teams for the 2016 season.
- Conduct a test of hypothesis to determine whether the mean salary of the teams was different from \$100.0 million. Use the .05 significance level.
  - Using a 5% significance level, conduct a test of hypothesis to determine whether the mean attendance was more than 2,000,000 per team.
46. **FILE** Refer to the Lincolnville School District bus data.
- Select the variable for the number of miles traveled last month. Conduct a hypothesis test to determine whether the mean miles traveled last month equals 10,000. Use the .01 significance level. Find the  $p$ -value and explain what it means.
  - A study of school bus fleets reports that the average maintenance cost per bus is \$4,000 per year. Using the maintenance cost variable, conduct a hypothesis test to determine whether the mean maintenance cost for Lincolnville’s bus fleet is more than \$4,000 at the .05 significance level. Determine the  $p$ -value and report the results.

## PRACTICE TEST

### Part 1—Objective

- The \_\_\_\_\_ is a statement about the value of a population parameter developed for the purpose of testing.
- We commit a Type II error when we \_\_\_\_\_ the null hypothesis when it is actually false.
- The probability of committing a Type I error is equal to the \_\_\_\_\_.
- The \_\_\_\_\_, based on sample information, is used to determine whether to reject the null hypothesis.
- The \_\_\_\_\_ value separates the region where the null hypothesis is rejected from the region where it is not rejected.
- In a \_\_\_\_\_-tailed test, the significance level is divided equally between the two tails. (one, two, neither)
- When conducting a test of hypothesis for means (assuming a normal population), we use the standard normal distribution when the population \_\_\_\_\_ is known.
- The \_\_\_\_\_ is the probability of finding a value of the test statistic at least as extreme as the one observed, given that the null hypothesis is true.
- The \_\_\_\_\_ conditions are necessary to conduct a test of hypothesis about a proportion.
- To conduct a test of proportions, the value of  $n(\pi)$  and  $n(1 - \pi)$  must be at least \_\_\_\_\_. (1, 5, 30, 1000)



**Part 2—Problems**

For each of these problems, use the six-step hypothesis-testing procedure.

1. The park manager at Fort Fisher State Park in North Carolina believes the typical park visitor spends at least 90 minutes in the park during the summer months. A sample of 18 visitors during the summer months of 2011 revealed the mean time in the park was 96 minutes with a standard deviation of 12 minutes. At the .01 significance level, is it reasonable to conclude that the mean time in the park is greater than 90 minutes?
2. The box fill weight of Frosted Flakes breakfast cereal follows the normal probability distribution with a mean of 9.75 ounces and a standard deviation of 0.27 ounces. A sample of 25 boxes filled this morning showed a mean of 9.85 ounces. Can we conclude that the mean weight is more than 9.75 ounces per box?
3. A recent newspaper article reported that for purchases of more than \$500, 67% of young married couples consulted with and sought the approval of their spouse. A sample of 300 young married couples in Chicago revealed 180 consulted with their spouse on their most recent purchase of more than \$500. At the .05 significance level, can we conclude that less than 67% of young married couples in Chicago sought the approval of their spouse?

# Two-Sample Tests of Hypothesis

# 11



©JGI/Blend Images LLC RF

- ▲ **GIBBS BABY FOOD COMPANY** wishes to compare the weight gain of infants using its brand versus its competitor's. A sample of 40 babies using the Gibbs products revealed a mean weight gain of 7.6 pounds in the first three months after birth. For the Gibbs brand, the population standard deviation of the sample is 2.3 pounds. A sample of 55 babies using the competitor's brand revealed a mean increase in weight of 8.1 pounds. The population standard deviation is 2.9 pounds. At the .05 significance level, can we conclude that babies using the Gibbs brand gained less weight? (See Exercise 3 and **LO11-1**.)

## LEARNING OBJECTIVES

---

*When you have completed this chapter, you will be able to:*

- LO11-1** Test a hypothesis that two independent population means are equal, assuming that the population standard deviations are known and equal.
- LO11-2** Test a hypothesis that two independent population means are equal, with unknown population standard deviations.
- LO11-3** Test a hypothesis about the mean population difference between paired or dependent observations.
- LO11-4** Explain the difference between dependent and independent samples.

## INTRODUCTION

Chapter 10 began our study of hypothesis testing. We described the nature of hypothesis testing and conducted tests of a hypothesis in which we compared the results of a single sample to a population value. That is, we selected a single random sample from a population and conducted a test of whether the proposed population value was reasonable. Recall in Chapter 10 that we selected a sample of the number of desks assembled per week at Jamestown Steel Company to determine whether there was a change



©Thinkstock/Getty Images

in the production rate. Similarly, we sampled the cost to process insurance claims to determine if cost-cutting measures resulted in a mean less than the current \$60 per claim. In both cases, we compared the results of a *single* sample statistic to a population parameter.

In this chapter, we expand the idea of hypothesis testing to two populations. That is, we select random samples from two different populations to determine whether the population means are equal. Some questions we might want to test are:

1. Is there a difference in the mean value of residential real estate sold by male agents and female agents in south Florida?
2. At Grabit Software Inc., do customer service employees receive more calls for assistance during the morning or afternoon?
3. In the fast-food industry, is there a difference in the mean number of days absent between young workers (under 21 years of age) and older workers (more than 60 years of age)?
4. Is there an increase in the production rate if music is piped into the production area?

We begin this chapter with the case in which we select random samples from two independent populations and wish to investigate whether these populations have the same mean.

### LO11-1

Test a hypothesis that two independent population means are equal, assuming that the population standard deviations are known and equal.

## TWO-SAMPLE TESTS OF HYPOTHESIS: INDEPENDENT SAMPLES

A city planner in Tampa, Florida, wishes to know whether there is a difference in the mean hourly wage rate of plumbers and electricians in central Florida. A financial accountant wishes to know whether the mean rate of return for domestic, U.S. mutual funds is different from the mean rate of return on global mutual funds. In each of these cases, there are two independent populations. In the first case, the plumbers represent one population and the electricians, the other. In the second case, domestic, U.S. mutual funds are one population and global mutual funds, the other.

To investigate the question in each of these cases, we would select a random sample from each population and compute the mean of the two samples. If the two population means are the same, that is, the mean hourly rate is the same for the plumbers and the electricians, we would expect the *difference* between the two sample means to be zero. But what if our sample results yield a difference other than zero? Is that difference due to chance or is it because there is a real difference in the hourly earnings? A two-sample test of means will help to answer this question.

Return to the results of Chapter 8. Recall that we showed that a distribution of sample means would tend to approximate the normal distribution. We need to again assume that a distribution of sample means will follow the normal distribution. It can be shown

mathematically that the distribution of the differences between sample means for two normal distributions is also normal.

We can illustrate this theory in terms of the city planner in Tampa, Florida. To begin, let's assume some information that is not usually available. Suppose that the population of plumbers has a mean of \$30.00 per hour and a standard deviation of \$5.00 per hour. The population of electricians has a mean of \$29.00 and a standard deviation of \$4.50. Now, from this information it is clear that the two population means are not the same. The plumbers actually earn \$1.00 per hour more than the electricians. But we cannot expect to uncover this difference each time we sample the two populations.

Suppose we select a random sample of 40 plumbers and a random sample of 35 electricians and compute the mean of each sample. Then, we determine the difference between the sample means. It is this difference between the sample means that holds our interest. If the populations have the same mean, then we would expect the difference between the two sample means to be zero. If there is a difference between the population means, then we expect to find a difference between the sample means.

To understand the theory, we need to take several pairs of samples, compute the mean of each, determine the difference between the sample means, and study the distribution of the differences in the sample means. Because of the central limit theorem in Chapter 8, we know that the distribution of the sample means follows the normal distribution. If the two distributions of sample means follow the normal distribution, then we can reason that the distribution of their differences will also follow the normal distribution. This is the first hurdle.

The second hurdle refers to the mean of this distribution of differences. If we find the mean of this distribution is zero, that implies that there is no difference in the two populations. On the other hand, if the mean of the distribution of differences is equal to some value other than zero, either positive or negative, then we conclude that the two populations do not have the same mean.

To report some concrete results, let's return to the city planner in Tampa, Florida. Table 11–1 shows the result of selecting 20 different samples of 40 plumbers and 35 electricians, computing the mean of each sample, and finding the difference

**TABLE 11–1** The Mean Hourly Earnings of 20 Random Samples of Plumbers and Electricians and the Differences between the Means

| Sample | Plumbers | Electricians | Difference |
|--------|----------|--------------|------------|
| 1      | \$29.80  | \$28.76      | \$1.04     |
| 2      | 30.32    | 29.40        | 0.92       |
| 3      | 30.57    | 29.94        | 0.63       |
| 4      | 30.04    | 28.93        | 1.11       |
| 5      | 30.09    | 29.78        | 0.31       |
| 6      | 30.02    | 28.66        | 1.36       |
| 7      | 29.60    | 29.13        | 0.47       |
| 8      | 29.63    | 29.42        | 0.21       |
| 9      | 30.17    | 29.29        | 0.88       |
| 10     | 30.81    | 29.75        | 1.06       |
| 11     | 30.09    | 28.05        | 2.04       |
| 12     | 29.35    | 29.07        | 0.28       |
| 13     | 29.42    | 28.79        | 0.63       |
| 14     | 29.78    | 29.54        | 0.24       |
| 15     | 29.60    | 29.60        | 0.00       |
| 16     | 30.60    | 30.19        | 0.41       |
| 17     | 30.79    | 28.65        | 2.14       |
| 18     | 29.14    | 29.95        | −0.81      |
| 19     | 29.91    | 28.75        | 1.16       |
| 20     | 28.74    | 29.21        | −0.47      |

between the two sample means. In the first case, the sample of 40 plumbers has a mean of \$29.80, and for the 35 electricians, the mean is \$28.76. The difference between the sample means is \$1.04. This process was repeated 19 more times. Observe that in 17 of the 20 cases, the differences are positive because the mean of the plumbers is larger than the mean of the electricians. In two cases, the differences are negative because the mean of the electricians is larger than the mean of the plumbers. In one case, the means are equal.

Our final hurdle is that we need to know something about the *variability* of the distribution of differences. To put it another way, what is the standard deviation of this distribution of differences? Statistical theory shows that when we have independent populations, as in this case, the distribution of the differences has a variance (standard deviation squared) equal to the sum of the two individual variances. This means that we can add the variances of the two sampling distributions. To put it another way, the variance of the difference in sample means ( $\bar{x}_1 - \bar{x}_2$ ) is equal to the sum of the variance for the plumbers and the variance for the electricians.

**VARIANCE OF THE DISTRIBUTION OF DIFFERENCES IN MEANS**

$$\sigma_{\bar{x}_1 - \bar{x}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad (11-1)$$

The term  $\sigma_{\bar{x}_1 - \bar{x}_2}^2$  looks complex but need not be difficult to interpret. The  $\sigma^2$  portion reminds us that it is a variance, and the subscript  $\bar{x}_1 - \bar{x}_2$  indicates that it is a distribution of differences in the sample means.

We can put this equation in a more usable form by taking the square root, so that we have the standard deviation or “standard error” of the distribution of differences. Finally, we standardize the distribution of the differences. The result is the following equation.

**TWO-SAMPLE TEST OF MEANS—KNOWN  $\sigma$**

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (11-2)$$

Before we present an example, let’s review the assumptions necessary for using formula (11-2).

- The two populations follow normal distributions.
- The two samples are unrelated, that is, independent.
- The standard deviations for both populations are known.

The following example shows the details of the test of hypothesis for two population means and shows how to interpret the results.

► **EXAMPLE**

Customers at the FoodTown Supermarket have a choice when paying for their groceries. They may check out and pay using the standard cashier-assisted checkout, or they may use the new Fast Lane procedure. In the standard procedure, a FoodTown employee scans each item and puts it on a short conveyor, where another employee puts it in a bag and then into the grocery cart. In the Fast Lane procedure, the customer scans each item, bags it, and places



©photocritical/Shutterstock

the bags in the cart him- or herself. The Fast Lane procedure is designed to reduce the time a customer spends in the checkout line.

The Fast Lane facility was recently installed at the Byrne Road FoodTown location. The store manager would like to know if the mean checkout time using the standard checkout method is longer than using the Fast Lane. She gathered the following sample information. The time is measured from when the customer enters the line until all his or her bags are in the cart. Hence, the time includes both waiting in line and checking out. What is the  $p$ -value?

| Customer Type | Sample Size | Sample Mean  | Population Standard Deviation |
|---------------|-------------|--------------|-------------------------------|
| Standard      | 50          | 5.50 minutes | 0.40 minute                   |
| Fast Lane     | 100         | 5.30 minutes | 0.30 minute                   |

### SOLUTION

We use the six-step hypothesis-testing procedure to investigate the question.

**Step 1: State the null hypothesis and the alternate hypothesis.** The null hypothesis is that the mean standard checkout time is less than or equal to the mean Fast Lane checkout time. In other words, the difference of 0.20 minute between the mean checkout time for the standard method and the mean checkout time for Fast Lane is due to chance. The alternate hypothesis is that the mean checkout time is larger for those using the standard method. We will let  $\mu_S$  refer to the mean checkout time for the population of standard customers and  $\mu_F$  the mean checkout time for the Fast Lane customers. The null and alternative hypotheses are:

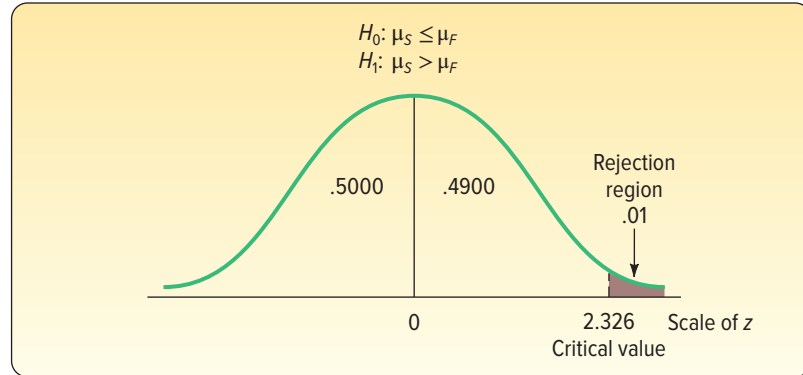
$$H_0: \mu_S \leq \mu_F$$

$$H_1: \mu_S > \mu_F$$

**Step 2: Select the level of significance.** The significance level is the probability that we reject the null hypothesis when it is actually true. This likelihood is determined prior to selecting the sample or performing any calculations. The .05 and .01 significance levels are the most common, but other values, such as .02 and .10, are also used. As a probability, the significance level must be between 0 and 1. However, the reasonable decision is to choose probabilities that are small. In this case, we selected the .01 significance level.

**Step 3: Determine the test statistic.** In Chapter 10, we used the standard normal distribution (that is,  $z$ ) and  $t$  as test statistics. In this case, we use the  $z$  distribution as the test statistic because we assume the two population distributions are both normal and the standard deviations of both populations are known.

**Step 4: Formulate a decision rule.** The decision rule is based on the null and the alternate hypotheses (i.e., one-tailed or two-tailed test), the level of significance, and the test statistic used. We selected the .01 significance level and the  $z$  distribution as the test statistic, and we wish to determine whether the mean checkout time is longer using the standard method. We set the alternate hypothesis to indicate that the mean checkout time is longer for those using the standard method than the Fast Lane method. Hence, the rejection region is in the upper tail of the standard normal distribution (a one-tailed test). To find the critical value, go to Student's  $t$  distribution



**CHART 11–1** Decision Rule for One-Tailed Test at .01 Significance Level

(Appendix B.5). In the table headings, find the row labeled “Level of Significance for One-Tailed Test” and select the column for an alpha of .01. Go to the bottom row with infinite degrees of freedom. The  $z$  critical value is 2.326, so the decision rule is to reject the null hypothesis if the value of the test statistic exceeds 2.326. Chart 11–1 depicts the decision rule.

**Step 5: Make the decision regarding  $H_0$ .** FoodTown randomly selected 50 customers using the standard checkout and computed a sample mean checkout time of 5.5 minutes, and 100 customers using the Fast Lane checkout and computed a sample mean checkout time of 5.3 minutes. We assume that the population standard deviations for the two methods are known. We use formula (11–2) to compute the value of the test statistic.

$$z = \frac{\bar{x}_S - \bar{x}_F}{\sqrt{\frac{\sigma_S^2}{n_S} + \frac{\sigma_F^2}{n_F}}} = \frac{5.5 - 5.3}{\sqrt{\frac{0.40^2}{50} + \frac{0.30^2}{100}}} = \frac{0.2}{0.064031} = 3.123$$

The computed value of 3.123 is larger than the critical value of 2.326. Our decision is to reject the null hypothesis and accept the alternate hypothesis.

**Step 6: Interpret the result.** The difference of .20 minute between the mean checkout times is too large to have occurred by chance. We conclude the Fast Lane method is faster.

What is the  $p$ -value for the test statistic? Recall that the  $p$ -value is the probability of finding a value of the test statistic this extreme when the null hypothesis is true. To calculate the  $p$ -value, we need the probability of a  $z$  value larger than 3.123. From Appendix B.3, we cannot find the probability associated with 3.123. The largest value available is 3.09. The area corresponding to 3.09 is .4990. In this case, we can report that the  $p$ -value is less than .0010, found by  $.5000 - .4990$ . We conclude that there is very little likelihood that the null hypothesis is true! The checkout time is less using the Fast Lane.

#### STATISTICS IN ACTION

Do you live to work or work to live? A recent poll of 802 working Americans revealed that, among those who considered their work as a career, the mean number of hours worked per day was 8.7. Among those who considered their work as a job, the mean number of hours worked per day was 7.6.

In summary, the criteria for using formula (11–2) are:

1. **The samples are from independent populations.** This means the checkout time for the Fast Lane customers is unrelated to the checkout time for the other customers. For example, Mr. Smith’s checkout time does not affect any other customer’s checkout time.

- Both populations follow the normal distribution.** In the FoodTown example, the population of times in both the standard checkout line and the Fast Lane follow normal distributions.
- Both population standard deviations are known.** In the FoodTown example, the population standard deviation of the Fast Lane times was 0.30 minute. The population standard deviation of the standard checkout times was 0.40 minute.

## SELF-REVIEW 11-1



Tom Sevits is the owner of the Appliance Patch. Recently Tom observed a difference in the dollar value of sales between the men and women he employs as sales associates. A sample of 40 days revealed the men sold a mean of \$1,400 worth of appliances per day. For a sample of 50 days, the women sold a mean of \$1,500 worth of appliances per day. Assume the population standard deviation for men is \$200 and for women \$250. At the .05 significance level, can Mr. Sevits conclude that the mean amount sold per day is larger for the women?

- State the null hypothesis and the alternate hypothesis.
- What is the decision rule?
- What is the value of the test statistic?
- What is your decision regarding the null hypothesis?
- What is the  $p$ -value?
- Interpret the result.

## EXERCISES

- A sample of 40 observations is selected from one population with a population standard deviation of 5. The sample mean is 102. A sample of 50 observations is selected from a second population with a population standard deviation of 6. The sample mean is 99. Conduct the following test of hypothesis using the .04 significance level.

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

- Is this a one-tailed or a two-tailed test?
  - State the decision rule.
  - Compute the value of the test statistic.
  - What is your decision regarding  $H_0$ ?
  - What is the  $p$ -value?
- A sample of 65 observations is selected from one population with a population standard deviation of 0.75. The sample mean is 2.67. A sample of 50 observations is selected from a second population with a population standard deviation of 0.66. The sample mean is 2.59. Conduct the following test of hypothesis using the .08 significance level.

$$H_0: \mu_1 \leq \mu_2$$

$$H_1: \mu_1 > \mu_2$$

- Is this a one-tailed or a two-tailed test?
- State the decision rule.
- Compute the value of the test statistic.
- What is your decision regarding  $H_0$ ?
- What is the  $p$ -value?

*Note:* Use the six-step hypothesis-testing procedure to solve the following exercises.

- Gibbs Baby Food Company wishes to compare the weight gain of infants using its brand versus its competitor's. A sample of 40 babies using the Gibbs products revealed a mean weight gain of 7.6 pounds in the first three months after birth. For the Gibbs brand, the population standard deviation of the sample is 2.3 pounds. A



sample of 55 babies using the competitor's brand revealed a mean increase in weight of 8.1 pounds. The population standard deviation is 2.9 pounds. At the .05 significance level, can we conclude that babies using the Gibbs brand gained less weight? Compute the  $p$ -value and interpret it.

4. As part of a study of corporate employees, the director of human resources for PNC Inc. wants to compare the distance traveled to work by employees at its office in downtown Cincinnati with the distance for those in downtown Pittsburgh. A sample of 35 Cincinnati employees showed they travel a mean of 370 miles per month. A sample of 40 Pittsburgh employees showed they travel a mean of 380 miles per month. The population standard deviations for the Cincinnati and Pittsburgh employees are 30 and 26 miles, respectively. At the .05 significance level, is there a difference in the mean number of miles traveled per month between Cincinnati and Pittsburgh employees?
5. Do married and unmarried women spend the same amount of time per week using Facebook? A random sample of 45 married women who use Facebook spent an average of 3.0 hours per week on this social media website. A random sample of 39 unmarried women who regularly use Facebook spent an average of 3.4 hours per week. Assume that the weekly Facebook time for married women has a population standard deviation of 1.2 hours, and the population standard deviation for unmarried Facebook users is 1.1 hours per week. Using the .05 significance level, do married and unmarried women differ in the amount of time per week spent on Facebook? Find the  $p$ -value and interpret the result.
6. Mary Jo Fitzpatrick is the vice president for Nursing Services at St. Luke's Memorial Hospital. Recently she noticed in the job postings for nurses that those that are unionized seem to offer higher wages. She decided to investigate and gathered the following information.

| Group    | Sample Size | Sample Mean Wage | Population Standard Deviation |
|----------|-------------|------------------|-------------------------------|
| Union    | 40          | \$20.75          | \$2.25                        |
| Nonunion | 45          | \$19.80          | \$1.90                        |

Would it be reasonable for her to conclude that union nurses earn more? Use the .02 significance level. What is the  $p$ -value?

### LO11-2

Test a hypothesis that two independent population means are equal, with unknown population standard deviations.

## COMPARING POPULATION MEANS WITH UNKNOWN POPULATION STANDARD DEVIATIONS

In the previous section, we used the standard normal distribution and  $z$  as the test statistic to test a hypothesis that two population means from independent populations were equal. The hypothesis tests presumed that the populations were normally distributed and that we knew the population standard deviations. However, in most cases, we do not know the population standard deviations. We can overcome this problem, as we did in the one-sample case in the previous chapter, by substituting the sample standard deviation ( $s$ ) for the population standard deviation ( $\sigma$ ). See formula (10–2) on page 290.

### Two-Sample Pooled Test

In this section, we describe another method for comparing the sample means of two independent populations to determine if the sampled populations could reasonably have the same mean. The method described does *not* require that we know the standard deviations of the populations. This gives us a great deal more flexibility when

investigating the difference between sample means. There are three major differences between this test and the previous test described in this chapter.

1. We assume the sampled populations have equal but unknown standard deviations.
2. We combine or “pool” the sample standard deviations.
3. We use the  $t$  distribution as the test statistic.

The formula for computing the value of the test statistic  $t$  is similar to formula (11–2), but an additional calculation is necessary. The two sample standard deviations are pooled to form a single estimate of the unknown population standard deviation. In essence, we compute a weighted mean of the two sample standard deviations and use this value as an estimate of the unknown population standard deviation. The weights are the degrees of freedom that each sample provides. Why do we need to pool the sample standard deviations? Because we assume that the two populations have equal standard deviations, the best estimate we can make of that value is to combine or pool all the sample information we have about the value of the population standard deviation.

The following formula is used to pool the sample standard deviations. Notice that two factors are involved: the number of observations in each sample and the sample standard deviations themselves.

$$\text{POOLED VARIANCE} \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (11-3)$$

where:

- $s_1^2$  is the variance (standard deviation squared) of the first sample.
- $s_2^2$  is the variance of the second sample.

The value of  $t$  is computed from the following equation.

$$\text{TWO-SAMPLE TEST OF MEANS—} \quad t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (11-4)$$

**UNKNOWN  $\sigma$ 'S**

where:

- $\bar{x}_1$  is the mean of the first sample.
- $\bar{x}_2$  is the mean of the second sample.
- $n_1$  is the number of observations in the first sample.
- $n_2$  is the number of observations in the second sample.
- $s_p^2$  is the pooled estimate of the population variance.

The number of degrees of freedom in the test is the total number of items sampled minus the total number of samples. Because there are two samples, there are  $n_1 + n_2 - 2$  degrees of freedom.

To summarize, there are three requirements or assumptions for the test.

1. The sampled populations are approximately normally distributed.
2. The sampled populations are independent.
3. The standard deviations of the two populations are equal.

The following example/solution explains the details of the test.

### EXAMPLE

Owens Lawn Care Inc. manufactures and assembles lawnmowers that are shipped to dealers throughout the United States and Canada. Two different procedures have been proposed for mounting the engine on the frame of the lawnmower. The question is: Is there a difference in the mean time to mount the engines on the

frames of the lawnmowers? The first procedure was developed by longtime Owens employee Herb Welles (designated as procedure W), and the other procedure was developed by Owens Vice President of Engineering William Atkins (designated as procedure A). To evaluate the two methods, we conduct a time and motion study. A sample of five employees is timed using the Welles method and six using the Atkins method. The results, in minutes, are shown below. Is there a difference in the mean mounting times? Use the .10 significance level.

| Welles<br>(minutes) | Atkins<br>(minutes) |
|---------------------|---------------------|
| 2                   | 3                   |
| 4                   | 7                   |
| 9                   | 5                   |
| 3                   | 8                   |
| 2                   | 4                   |
|                     | 3                   |

### SOLUTION

Following the six steps to test a hypothesis, the null hypothesis states that there is no difference in mean mounting times between the two procedures. The alternate hypothesis indicates that there is a difference.

$$H_0: \mu_W = \mu_A$$

$$H_1: \mu_W \neq \mu_A$$

The required assumptions are:

- The observations in the Welles sample are *independent* of the observations in the Atkins sample.
- The two populations follow the normal distribution.
- The two populations have equal standard deviations, but these standard deviations are not known.

Is there a difference between the mean assembly times using the Welles and the Atkins methods? The degrees of freedom are equal to the total number of items sampled minus the number of samples. In this case, that is  $n_W + n_A - 2$ . Five assemblers used the Welles method and six the Atkins method. Thus, there are 9 degrees of freedom, found by  $5 + 6 - 2$ . The critical values of  $t$ , from Appendix B.5 for  $df = 9$ , a two-tailed test, and the .10 significance level, are  $-1.833$  and  $1.833$ . The decision rule is portrayed graphically in Chart 11–2. We do not reject the null hypothesis if the computed value of  $t$  falls between  $-1.833$  and  $1.833$ .

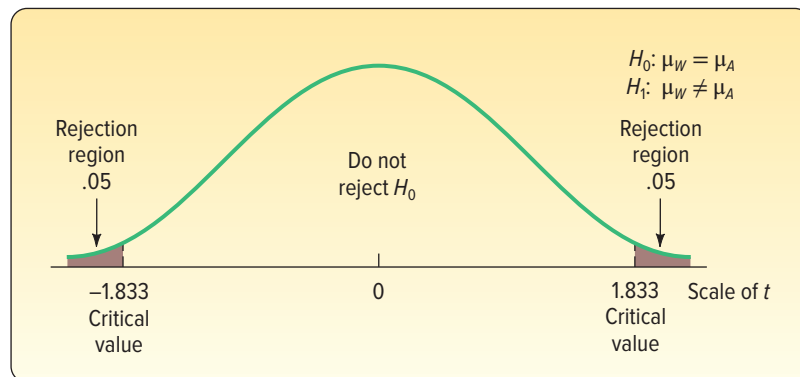


CHART 11–2 Regions of Rejection, Two-Tailed Test,  $df = 9$ , and .10 Significance Level

We use three steps to compute the value of  $t$ .

**Step 1: Calculate the sample standard deviations.** To compute the sample standard deviations, we use formula (3–8). See the details below.

| Welles Method   |                       | Atkins Method   |                            |
|-----------------|-----------------------|-----------------|----------------------------|
| $x_W$           | $(x_W - \bar{x}_W)^2$ | $x_A$           | $(x_A - \bar{x}_A)^2$      |
| 2               | $(2 - 4)^2 = 4$       | 3               | $(3 - 5)^2 = 4$            |
| 4               | $(4 - 4)^2 = 0$       | 7               | $(7 - 5)^2 = 4$            |
| 9               | $(9 - 4)^2 = 25$      | 5               | $(5 - 5)^2 = 0$            |
| 3               | $(3 - 4)^2 = 1$       | 8               | $(8 - 5)^2 = 9$            |
| 2               | $(2 - 4)^2 = 4$       | 4               | $(4 - 5)^2 = 1$            |
| $\overline{20}$ | $\overline{34}$       | $\overline{3}$  | $\overline{(3 - 5)^2 = 4}$ |
|                 |                       | $\overline{30}$ | $\overline{22}$            |

$$\bar{x}_W = \frac{\sum x_W}{n_W} = \frac{20}{5} = 4 \qquad \bar{x}_A = \frac{\sum x_A}{n_A} = \frac{30}{6} = 5$$

$$s_W = \sqrt{\frac{\sum (x_W - \bar{x}_W)^2}{n_W - 1}} = \sqrt{\frac{34}{5 - 1}} = 2.9155 \qquad s_A = \sqrt{\frac{\sum (x_A - \bar{x}_A)^2}{n_A - 1}} = \sqrt{\frac{22}{6 - 1}} = 2.0976$$

**Step 2: Pool the sample variances.** We use formula (11–3) to pool the sample variances (standard deviations squared).

$$s_p^2 = \frac{(n_W - 1)s_W^2 + (n_A - 1)s_A^2}{n_W + n_A - 2} = \frac{(5 - 1)(2.9155)^2 + (6 - 1)(2.0976)^2}{5 + 6 - 2} = 6.2222$$

**Step 3: Determine the value of  $t$ .** The mean mounting time for the Welles method is 4.00 minutes, found by  $\bar{x}_W = 20/5$ . The mean mounting time for the Atkins method is 5.00 minutes, found by  $\bar{x}_A = 30/6$ . We use formula (11–4) to calculate the value of  $t$ .

$$t = \frac{\bar{x}_W - \bar{x}_A}{\sqrt{s_p^2 \left( \frac{1}{n_W} + \frac{1}{n_A} \right)}} = \frac{4.00 - 5.00}{\sqrt{6.2222 \left( \frac{1}{5} + \frac{1}{6} \right)}} = -0.662$$

The decision is not to reject the null hypothesis because  $-0.662$  falls in the region between  $-1.833$  and  $1.833$ . Our conclusion is that the sample data failed to show a difference between the mean assembly times of the two methods.

We also can estimate the  $p$ -value using Appendix B.5. Locate the row with 9 degrees of freedom, and use the two-tailed test column. Find the  $t$  value, without regard to the sign, that is closest to our computed value of  $0.662$ . It is  $1.383$ , corresponding to a significance level of  $.20$ . Thus, even had we used the  $20\%$  significance level, we would not have rejected the null hypothesis of equal means. We can report that the  $p$ -value is greater than  $.20$ .

---

Excel has a procedure called “t-Test: Two Sample Assuming Equal Variances” that will perform the calculations of formulas (11–3) and (11–4) as well as find the sample means and sample variances. The details of the procedure are provided in Appendix C. The data are input in the first two columns of the Excel spreadsheet. They are labeled “Welles” and “Atkins.” The output follows. The value of  $t$ , called the “t Stat,” is  $-0.662$ , and the two-tailed  $p$ -value is  $.525$ . As we would expect, the exact  $p$ -value is larger than the significance level of  $.10$ . The conclusion is not to reject the null hypothesis.

|    | A      | B      | C | D   | E      | F      |
|----|--------|--------|---|---|--------|--------|
| 1  | Welles | Atkins |   | t-Test: Two-Sample Assuming Equal Variances |        |        |
| 2  | 2      | 3      |   |   |        |        |
| 3  | 4      | 7      |   |   | Welles | Atkins |
| 4  | 9      | 5      |   | Mean  | 4.000  | 5.000  |
| 5  | 3      | 8      |   | Variance                                    | 8.500  | 4.400  |
| 6  | 2      | 4      |   | Observations                                | 5.000  | 6.000  |
| 7  |        | 3      |   | Pooled Variance                             | 6.222  |        |
| 8  |        |        |   | Hypothesized Mean Difference                | 0.000  |        |
| 9  |        |        |   | df  | 9.000  |        |
| 10 |        |        |   | t Stat                                      | -0.662 |        |
| 11 |        |        |   | P(T<=t) one-tail                            | 0.262  |        |
| 12 |        |        |   | t Critical one-tail                         | 1.833  |        |
| 13 |        |        |   | P(T<=t) two-tail                            | 0.525  |        |
| 14 |        |        |   | t Critical two-tail                         | 2.262  |        |

Source: Microsoft Excel

## SELF-REVIEW 11-2



The production manager at Bellevue Steel, a manufacturer of wheelchairs, wants to compare the number of defective wheelchairs produced on the day shift with the number on the afternoon shift. A sample of the production from 6 day shifts and 8 afternoon shifts revealed the following number of defects.

|                  |   |    |   |    |   |    |    |   |
|------------------|---|----|---|----|---|----|----|---|
| <b>Day</b>       | 5 | 8  | 7 | 6  | 9 | 7  |    |   |
| <b>Afternoon</b> | 8 | 10 | 7 | 11 | 9 | 12 | 14 | 9 |

At the .05 significance level, is there a difference in the mean number of defects per shift?

- State the null hypothesis and the alternate hypothesis.
- What is the decision rule?
- What is the value of the test statistic?
- What is your decision regarding the null hypothesis?
- What is the  $p$ -value?
- Interpret the result.
- What are the assumptions necessary for this test?

## EXERCISES

For Exercises 7 and 8: (a) state the decision rule, (b) compute the pooled estimate of the population variance, (c) compute the test statistic, (d) state your decision about the null hypothesis, and (e) estimate the  $p$ -value.

7. The null and alternate hypotheses are:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

A random sample of 10 observations from one population revealed a sample mean of 23 and a sample standard deviation of 4. A random sample of 8 observations from another population revealed a sample mean of 26 and a sample standard deviation of 5. At the .05 significance level, is there a difference between the population means?

8. The null and alternate hypotheses are:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

A random sample of 15 observations from the first population revealed a sample mean of 350 and a sample standard deviation of 12. A random sample of 17 observations from the second population revealed a sample mean of 342 and a sample standard deviation of 15. At the .10 significance level, is there a difference in the population means?

Note: Use the six-step hypothesis testing procedure for the following exercises.

9. **FILE** Listed below are the 25 players on the opening-day roster of the 2016 New York Yankees Major League Baseball team, their salaries, and fielding positions.

| Player          | Position          | Salary (US\$) |
|-----------------|-------------------|---------------|
| C.C. Sabathia   | Starting Pitcher  | \$25,000,000  |
| Mark Teixeira   | First Base        | \$23,125,000  |
| Masahiro Tanaka | Starting Pitcher  | \$22,000,000  |
| Jacoby Ellsbury | Center Field      | \$21,142,857  |
| Alex Rodriguez  | Designated Hitter | \$21,000,000  |
| Brian McCann    | Catcher           | \$17,000,000  |
| Carlos Beltran  | Right Field       | \$15,000,000  |
| Brett Gardner   | Left Field        | \$13,500,000  |
| Chase Headley   | Third Base        | \$13,000,000  |
| Andrew Miller   | Relief Pitcher    | \$ 9,000,000  |
| Starlin Castro  | Second Base       | \$ 7,857,142  |
| Nathan Eovaldi  | Starting Pitcher  | \$ 5,600,000  |
| Michael Pineda  | Starting Pitcher  | \$ 4,300,000  |
| Ivan Nova       | Relief Pitcher    | \$ 4,100,000  |
| Dustin Ackley   | Left Field        | \$ 3,200,000  |
| Didi Gregorius  | Shortstop         | \$ 2,425,000  |
| Aaron Hicks     | Center Field      | \$ 574,000    |
| Austin Romine   | Catcher           | \$ 556,000    |
| Chasen Shreve   | Relief Pitcher    | \$ 533,400    |
| Luis Severino   | Starting Pitcher  | \$ 521,300    |
| Kirby Yates     | Relief Pitcher    | \$ 511,900    |
| Ronald Torreyes | Second Base       | \$ 508,600    |
| Johnny Barbato  | Relief Pitcher    | \$ 507,500    |
| Dellin Betances | Relief Pitcher    | \$ 507,500    |
| Luis Cessa      | Relief Pitcher    | \$ 507,500    |

Sort the players into two groups, all pitchers (relief and starting) and position players (all others). Assume equal population standard deviations for the pitchers and the position players. Test the hypothesis that mean salaries of pitchers and position players are equal, using the .01 significance level.

10. A recent study compared the time spent together by single- and dual-earner couples. According to the records kept by the wives during the study, the mean amount of time spent together watching television among the single-earner couples was 61 minutes per day, with a standard deviation of 15.5 minutes. For the dual-earner couples, the mean number of minutes spent watching television was 48.4 minutes, with a standard deviation of 18.1 minutes. At the .01 significance level, can we conclude that the single-earner couples on average spend more time watching television together? There were 15 single-earner and 12 dual-earner couples studied.

11. **FILE** Ms. Lisa Monnin is the budget director for Nexus Media Inc. She would like to compare the daily travel expenses for the sales staff and the audit staff. She collected the following sample information.

|                   |     |     |     |     |     |     |     |
|-------------------|-----|-----|-----|-----|-----|-----|-----|
| <b>Sales (\$)</b> | 131 | 135 | 146 | 165 | 136 | 142 |     |
| <b>Audit (\$)</b> | 130 | 102 | 129 | 143 | 149 | 120 | 139 |

At the .10 significance level, can she conclude that the mean daily expenses are greater for the sales staff than the audit staff? What is the  $p$ -value?

12. **FILE** The Tampa Bay (Florida) Area Chamber of Commerce wanted to know whether the mean weekly salary of nurses was larger than that of schoolteachers. To investigate, they collected the following information on the amounts earned last week by a sample of schoolteachers and a sample of nurses.

|                            |       |       |       |       |       |       |       |       |       |       |       |       |
|----------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| <b>Schoolteachers (\$)</b> | 1,095 | 1,076 | 1,077 | 1,125 | 1,034 | 1,059 | 1,052 | 1,070 | 1,079 | 1,080 | 1,092 | 1,082 |
| <b>Nurses (\$)</b>         | 1,091 | 1,140 | 1,071 | 1,021 | 1,100 | 1,109 | 1,075 | 1,079 |       |       |       |       |

Is it reasonable to conclude that the mean weekly salary of nurses is higher? Use the .01 significance level. What is the  $p$ -value?

**LO11-3**

Test a hypothesis about the mean population difference between paired or dependent observations.

## TWO-SAMPLE TESTS OF HYPOTHESIS: DEPENDENT SAMPLES

In the Owens Lawn Care example/solution on page 313, we tested the difference between the means from two independent populations. We compared the mean time required to mount an engine using the Welles method to the time to mount the engine using the Atkins method. The samples were *independent*, meaning that the sample of assembly times using the Welles method was in no way related to the sample of assembly times using the Atkins method.

There are situations, however, in which the samples are not independent. To put it another way, the samples are *dependent* or *related*. As an example, Nickel Savings and Loan employs two firms, Schadek Appraisals and Bowyer Real Estate, to appraise the value of the real estate properties on which it makes loans. It is important that these two firms be similar in their appraisal values. To review the consistency of the two appraisal firms, Nickel Savings randomly selects 10 homes and has both Schadek Appraisals and Bowyer Real Estate appraise the values of the selected homes. For each home, there will be a pair of appraisal values. That is, for each home there will be an appraised value from both Schadek Appraisals and Bowyer Real Estate. The appraised values depend on, or are related to, the home selected. This is also referred to as a **paired sample**.

For hypothesis testing, we are interested in the distribution of the *differences* in the appraised value of each home. Hence, there is only one sample. To put it more formally, we are investigating whether the mean of the distribution of differences in the appraised values is 0. The sample is made up of the *differences* between the appraised values determined by Schadek Appraisals and the values from Bowyer Real Estate. If the two appraisal firms are reporting similar estimates, then sometimes Schadek Appraisals will have the higher value and sometimes Bowyer Real Estate will have the higher value. However, the mean of the distribution of differences will be 0. On the other hand, if



©David Buffington/Getty Images

one of the firms consistently reports larger appraisal values, then the mean of the distribution of the differences will not be 0.

We will use the symbol  $\mu_d$  to indicate the population mean of the distribution of differences. We assume the distribution of the population of differences is approximately normally distributed. The test statistic follows the  $t$  distribution, and we calculate its value from the following formula:

**PAIRED  $t$  TEST**

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

**(11-5)**

There are  $n - 1$  degrees of freedom and

$\bar{d}$  is the mean of the differences between the paired or related observations.

$s_d$  is the standard deviation of the differences between the paired or related observations.

$n$  is the number of paired observations.

The standard deviation of the differences is computed by the familiar formula for the standard deviation [see formula (3-8)], except  $d$  is substituted for  $x$ . The formula is:

$$s_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n - 1}}$$

The following example illustrates this test.

**EXAMPLE**

Recall that Nickel Savings and Loan wishes to compare the two companies it uses to appraise the value of residential homes. Nickel Savings selected a sample of 10 residential properties and scheduled both firms for an appraisal. The results, reported in \$000, are:

| Home | Schadek | Bowyer |
|------|---------|--------|
| A    | 235     | 228    |
| B    | 210     | 205    |
| C    | 231     | 219    |
| D    | 242     | 240    |
| E    | 205     | 198    |
| F    | 230     | 223    |
| G    | 231     | 227    |
| H    | 210     | 215    |
| I    | 225     | 222    |
| J    | 249     | 245    |

At the .05 significance level, can we conclude there is a difference between the firms' mean appraised home values?

**SOLUTION**

The first step is to state the null and the alternate hypotheses. In this case, a two-tailed alternative is appropriate because we are interested in determining whether



there is a *difference* in the firms' appraised values. We are not interested in showing whether one particular firm appraises property at a higher value than the other. The question is whether the sample differences in the appraised values could have come from a population with a mean of 0. If the population mean of the differences is 0, then we conclude that there is no difference between the two firms' appraised values. The null and alternate hypotheses are:

$$H_0: \mu_d = 0$$

$$H_1: \mu_d \neq 0$$

There are 10 homes appraised by both firms, so  $n = 10$ , and  $df = n - 1 = 10 - 1 = 9$ . We have a two-tailed test, and the significance level is .05. To determine the critical value, go to Appendix B.5 and move across the row with 9 degrees of freedom to the column for a two-tailed test and the .05 significance level. The value at the intersection is 2.262. This value appears in the box in Table 11–2. The decision rule is to reject the null hypothesis if the computed value of  $t$  is less than  $-2.262$  or greater than 2.262. Here are the computational details.

| Home | Schadek | Bowyer | Difference, $d$ | $(d - \bar{d})$ | $(d - \bar{d})^2$ |
|------|---------|--------|-----------------|-----------------|-------------------|
| A    | 235     | 228    | 7               | 2.4             | 5.76              |
| B    | 210     | 205    | 5               | 0.4             | 0.16              |
| C    | 231     | 219    | 12              | 7.4             | 54.76             |
| D    | 242     | 240    | 2               | -2.6            | 6.76              |
| E    | 205     | 198    | 7               | 2.4             | 5.76              |
| F    | 230     | 223    | 7               | 2.4             | 5.76              |
| G    | 231     | 227    | 4               | -0.6            | 0.36              |
| H    | 210     | 215    | -5              | -9.6            | 92.16             |
| I    | 225     | 222    | 3               | -1.6            | 2.56              |
| J    | 249     | 245    | 4               | -0.6            | 0.36              |
|      |         |        | 46              | 0               | 174.40            |

$$\bar{d} = \frac{\Sigma d}{n} = \frac{46}{10} = 4.60$$

$$s_d = \sqrt{\frac{\Sigma(d - \bar{d})^2}{n - 1}} = \sqrt{\frac{174.4}{10 - 1}} = 4.402$$

Using formula (11–5), the value of the test statistic is 3.305, found by

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{4.6}{4.402 / \sqrt{10}} = \frac{4.6}{1.3920} = 3.305$$

Because the computed  $t$  falls in the rejection region, the null hypothesis is rejected. The population distribution of differences does not have a mean of 0. We conclude that there is a difference between the firms' mean appraised home values. The largest difference of \$12,000 is for Home 3. Perhaps that would be an appropriate place to begin a more detailed review.

To find the  $p$ -value, we use Appendix B.5 and the section for a two-tailed test. Move along the row with 9 degrees of freedom and find the values of  $t$  that are closest to our calculated value. For a .01 significance level, the value of  $t$  is 3.250. The computed value is larger than this value, but smaller than the value of 4.781 corresponding to the .001 significance level. Hence, the  $p$ -value is between .01 and .001. This information is highlighted in Table 11–2.

**TABLE 11-2** A Portion of the *t* Distribution from Appendix B.5

| Confidence Intervals |   |       |        |        |        |  |
|----------------------|---|-------|--------|--------|--------|--|
|                      | 80%                                       | 90%   | 95%    | 98%    | 99%    | <i>p</i> -value between 0.01 and 0.001 |
| <i>df</i>            | Level of Significance for One-Tailed Test |       |        |        |        |  |
|                      | 0.10                                      | 0.05  | 0.025  | 0.01   | 0.005  |  |
|                      | Level of Significance for Two-Tailed Test |       |        |        |        |  |
|                      | 0.20                                      | 0.10  | 0.05   | 0.02   | 0.01   | 0.001                                  |
| 1                    | 3.078                                     | 6.314 | 12.706 | 31.821 | 63.657 | 636.619                                |
| 2                    | 1.886                                     | 2.920 | 4.303  | 6.965  | 9.925  | 31.599                                 |
| 3                    | 1.638                                     | 2.353 | 3.182  | 4.541  | 5.841  | 12.924                                 |
|                      |   | 2.132 | 2.776  | 3.747  | 4.604  | 8.610                                  |
|                      |   | 2.015 | 2.571  | 3.365  | 4.032  | 6.869                                  |
| 6                    | 1.440                                     | 1.943 | 2.447  | 3.143  | 3.707  | 5.959                                  |
| 7                    | 1.415                                     | 1.895 | 2.365  | 2.998  | 3.499  | 5.408                                  |
| 8                    | 1.397                                     | 1.860 | 2.306  | 2.896  | 3.355  | 5.041                                  |
| 9                    | 1.383                                     | 1.833 | 2.262  | 2.821  | 3.250  | 4.781                                  |
| 10                   | 1.372                                     | 1.812 | 2.228  | 2.764  | 3.169  | 4.587                                  |

Excel’s statistical analysis software has a procedure called “t-Test: Paired Two-Sample for Means” that will perform the calculations of formula (11–5). The output from this procedure follows.

The computed value of *t* is 3.305, and the two-tailed *p*-value is .009. Because the *p*-value is less than .05, we reject the hypothesis that the mean of the distribution of the differences between the appraised values is zero.

|    | A    | B       | C      | D | E                                   | F       | G       |
|----|------|---------|--------|---|-------------------------------------|---------|---------|
| 1  | Home | Schadek | Bowyer |   | t-Test: Paired Two-Sample for Means |         |         |
| 2  | A    | 235     | 228    |   |                                     |         |         |
| 3  | B    | 210     | 205    |   |                                     | Schadek | Bowyer  |
| 4  | C    | 231     | 219    |   | Mean                                | 226.800 | 222.200 |
| 5  | D    | 242     | 240    |   | Variance                            | 208.844 | 204.178 |
| 6  | E    | 205     | 198    |   | Observations                        | 10.000  | 10.000  |
| 7  | F    | 230     | 223    |   | Pearson Correlation                 | 0.953   |         |
| 8  | G    | 231     | 227    |   | Hypothesized Mean Difference        | 0.000   |         |
| 9  | H    | 210     | 215    |   | df                                  | 9.000   |         |
| 10 | I    | 225     | 222    |   | t Stat                              | 3.305   |         |
| 11 | J    | 249     | 245    |   | P(T<=t) one-tail                    | 0.005   |         |
| 12 |      |         |        |   | t Critical one-tail                 | 1.833   |         |
| 13 |      |         |        |   | P(T<=t) two-tail                    | 0.009   |         |
| 14 |      |         |        |   | t Critical two-tail                 | 2.262   |         |
| 15 |      |         |        |   |                                     |         |         |

Source: Microsoft Excel

**LO11-4**

Explain the difference between dependent and independent samples.

## COMPARING DEPENDENT AND INDEPENDENT SAMPLES

Beginning students are often confused by the difference between tests for independent samples [formula (11–4)] and tests for dependent samples [formula (11–5)]. How do we tell the difference between dependent and independent samples? There are two types of dependent samples: (1) those characterized by a measurement, an intervention of some type, and then another measurement; and (2) a matching or pairing of the observations. To explain further:

1. The first type of dependent sample is characterized by a measurement followed by an intervention of some kind and then another measurement. This could be called a

“before” and “after” study. Two examples will help to clarify. Suppose we want to show that, by placing speakers in the production area and playing soothing music, we are able to increase production. We begin by selecting a sample of workers and measuring their output under the current conditions. The speakers are then installed in the production area, and we again measure the output of the same workers. There are two measurements, before placing the speakers in the production area and after. The intervention is placing speakers in the production area. A crucial factor of dependent sampling is that each production worker is measured before and after installing the speakers.

A second example involves an educational firm that offers courses designed to increase test scores and reading ability. Suppose the firm wants to offer a course that will help high school juniors increase their SAT scores. To begin, each student takes the SAT in the junior year in high school. During the summer between the junior and senior year, they participate in the course that gives them tips on taking tests. Finally, during the fall of their senior year in high school, they retake the SAT. Again, the procedure is characterized by a measurement (taking the SAT as a junior), an intervention (the summer workshops), and another measurement (taking the SAT during their senior year).

- The second type of dependent sample is characterized by matching or pairing observations. The previous example/solution regarding Nickel Savings illustrates dependent samples. A property is selected and both firms appraise the same property. As a second example, suppose an industrial psychologist wishes to study the intellectual similarities of newly married couples. She selects a sample of newlyweds. Next, she administers a standard intelligence test to both the man and woman to determine the difference in the scores. Notice the matching that occurred: comparing the scores that are paired or matched by marriage.

Why do we prefer dependent samples to independent samples? By using dependent samples, we are able to reduce the variation in the sampling distribution. To illustrate, we will use the Nickel Savings and Loan example/solution just completed. Suppose we inappropriately decide that we have two independent samples of real estate property for appraisal and conduct the following test of hypothesis, using formula (11–4). The null and alternate hypotheses are:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

There are now two independent samples of 10 each. So the number of degrees of freedom is  $10 + 10 - 2 = 18$ . From Appendix B.5, for the .05 significance level,  $H_0$  is rejected if  $t$  is less than  $-2.101$  or greater than  $2.101$ .

We use Excel to find the means and standard deviations of the two independent samples as shown in the Chapter 3 section of Appendix C. The Excel instructions to find the pooled variance and the value of the “t Stat” are in the Chapter 11 section in Appendix C. These values are highlighted in yellow.

|    | A      | B       | C      | D | E   | F       | G       | H |
|----|--------|---------|--------|---|---|---------|---------|---|
| 1  | Home   | Schadek | Bowyer |   | t-Test: Two-Sample Assuming Equal Variances |         |         |   |
| 2  | A      | 235     | 228    |   |   |         |         |   |
| 3  | B      | 210     | 205    |   |   | Schadek | Bowyer  |   |
| 4  | C      | 231     | 219    |   | Mean  | 226.800 | 222.200 |   |
| 5  | D      | 242     | 240    |   | Variance                                    | 208.844 | 204.178 |   |
| 6  | E      | 205     | 198    |   | Observations                                | 10.000  | 10.000  |   |
| 7  | F      | 230     | 223    |   | Pooled Variance                             | 206.511 |         |   |
| 8  | G      | 231     | 227    |   | Hypothesized Mean Difference                | 0.000   |         |   |
| 9  | H      | 210     | 215    |   | df  | 18.000  |         |   |
| 10 | I      | 225     | 222    |   | t Stat                                      | 0.716   |         |   |
| 11 | J      | 249     | 245    |   | P(T<=t) one-tail                            | 0.242   |         |   |
| 12 |        |         |        |   | t Critical one-tail                         | 1.734   |         |   |
| 13 | Mean = | 226.80  | 222.20 |   | P(T<=t) two-tail                            | 0.483   |         |   |
| 14 | S =    | 14.45   | 14.29  |   | t Critical two-tail                         | 2.101   |         |   |
| 15 |        |         |        |   |   |         |         |   |

Source: Microsoft Excel

The mean of the appraised value of the 10 properties by Schadek is \$226,800, and the standard deviation is \$14,500. For Bowyer Real Estate, the mean appraised value is \$222,200, and the standard deviation is \$14,290. To make the calculations easier, we use \$000 instead of \$. The value of the pooled estimate of the variance from formula (11-3) is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(10 - 1)(14.45^2) + (10 - 1)(14.29)^2}{10 + 10 - 2} = 206.50$$

From formula (11-4),  $t$  is 0.716.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{226.8 - 222.2}{\sqrt{206.50 \left( \frac{1}{10} + \frac{1}{10} \right)}} = \frac{4.6}{6.4265} = 0.716$$

The computed  $t$  (0.716) is less than 2.101, so the null hypothesis is not rejected. We cannot show that there is a difference in the mean appraisal value. That is not the same conclusion that we got before! Why does this happen? The numerator is the same in the paired observations test (4.6). However, the denominator is smaller. In the paired test, the denominator is 1.3920 (see the calculations on page 320 in the previous section). In the case of the independent samples, the denominator is 6.4265. There is more variation or uncertainty. This accounts for the difference in the  $t$  values and the difference in the statistical decisions. The denominator measures the standard error of the statistic. When the samples are *not* paired, two kinds of variation are present: differences between the two appraisal firms and the difference in the value of the real estate. Properties numbered 4 and 10 have relatively high values, whereas number 5 is relatively low. These data show how different the values of the property are, but we are really interested in the difference between the two appraisal firms.

In sum, when we can pair or match observations that measure differences for a common variable, a hypothesis test based on dependent samples is more sensitive to detecting a significant difference than a hypothesis test based on independent samples. In the case of comparing the property valuations by Schadek Appraisals and Bowyer Real Estate, the hypothesis test based on dependent samples eliminates the variation between the values of the properties and focuses only on the comparisons in the two appraisals for each property. There is a bit of bad news here. In the dependent samples test, the degrees of freedom are half of what they are if the samples are not paired. For the real estate example, the degrees of freedom drop from 18 to 9 when the observations are paired. However, in most cases, this is a small price to pay for a better test.

## SELF-REVIEW 11-3



Advertisements by Core Fitness Center claim that completing its course will result in losing weight. A random sample of eight recent participants showed the following weights before and after completing the course. At the .01 significance level, can we conclude the students lost weight?

| Name     | Before | After |
|----------|--------|-------|
| Hunter   | 155    | 154   |
| Cashman  | 228    | 207   |
| Mervine  | 141    | 147   |
| Massa    | 162    | 157   |
| Creola   | 211    | 196   |
| Peterson | 164    | 150   |
| Redding  | 184    | 170   |
| Poust    | 172    | 165   |

- State the null hypothesis and the alternate hypothesis.
- What is the critical value of  $t$ ?
- What is the computed value of  $t$ ?
- Interpret the result. What is the  $p$ -value?
- What assumption needs to be made about the distribution of the differences?

## EXERCISES

13. The null and alternate hypotheses are:

$$H_0: \mu_d \leq 0$$

$$H_1: \mu_d > 0$$

The following sample information shows the number of defective units produced on the day shift and the afternoon shift for a sample of four days last month.

|                 | Day |    |    |    |
|-----------------|-----|----|----|----|
|                 | 1   | 2  | 3  | 4  |
| Day shift       | 10  | 12 | 15 | 19 |
| Afternoon shift | 8   | 9  | 12 | 15 |

At the .05 significance level, can we conclude there are more defects produced on the day shift?

14. The null and alternate hypotheses are:

$$H_0: \mu_d = 0$$

$$H_1: \mu_d \neq 0$$

The following paired observations show the number of traffic citations given for speeding by Officer Dhondt and Officer Meredith of the South Carolina Highway Patrol for the last five months.

|                  | Number of Citations Issued |      |      |        |           |
|------------------|----------------------------|------|------|--------|-----------|
|                  | May                        | June | July | August | September |
| Officer Dhondt   | 30                         | 22   | 25   | 19     | 26        |
| Officer Meredith | 26                         | 19   | 20   | 15     | 19        |

At the .05 significance level, is there a difference in the mean number of citations given by the two officers?

*Note:* Use the six-step hypothesis testing procedure to solve the following exercises.

15. **FILE** The management of Discount Furniture, a chain of discount furniture stores in the Northeast, designed an incentive plan for salespeople. To evaluate this innovative plan, 12 salespeople were selected at random, and their weekly incomes before and after the plan were recorded.

| Salesperson   | Before | After |
|---------------|--------|-------|
| Sid Mahone    | \$320  | \$340 |
| Carol Quick   | 290    | 285   |
| Tom Jackson   | 421    | 475   |
| Andy Jones    | 510    | 510   |
| Jean Sloan    | 210    | 210   |
| Jack Walker   | 402    | 500   |
| Peg Mancuso   | 625    | 631   |
| Anita Loma    | 560    | 560   |
| John Cuso     | 360    | 365   |
| Carl Utz      | 431    | 431   |
| A. S. Kushner | 506    | 525   |
| Fern Lawton   | 505    | 619   |

Was there a significant increase in the typical salesperson's weekly income due to the innovative incentive plan? Use the .05 significance level. Estimate the  $p$ -value, and interpret it.

16. **FILE** The federal government recently granted funds for a special program designed to reduce crime in high-crime areas. A study of the results of the program in eight high-crime areas of Miami, Florida, yielded the following results.

|        |  | Number of Crimes by Area |   |   |   |    |    |   |   |
|--------|--|--------------------------|---|---|---|----|----|---|---|
|        |  | A                        | B | C | D | E  | F  | G | H |
| Before |  | 14                       | 7 | 4 | 5 | 17 | 12 | 8 | 9 |
| After  |  | 2                        | 7 | 3 | 6 | 8  | 13 | 3 | 5 |

Has there been a decrease in the number of crimes since the inauguration of the program? Use the .01 significance level. Estimate the  $p$ -value.

## CHAPTER SUMMARY

- I. In comparing two population means, we wish to know whether they could be equal.
- A. We are investigating whether the distribution of the difference between the means could have a mean of 0.
  - B. The test statistic follows the standard normal distribution if the population standard deviations are known.
    1. The two populations follow normal distributions.
    2. The samples are from independent populations.
    3. The formula to compute the value of  $z$  is

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (11-2)$$

- II. The test statistic to compare two means is the  $t$  distribution if the population standard deviations are not known.
- A. Both populations are approximately normally distributed.
  - B. The populations must have equal standard deviations.
  - C. The samples are independent.
  - D. Finding the value of  $t$  requires two steps.
    1. The first step is to pool the standard deviations according to the following formula:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (11-3)$$

2. The value of  $t$  is computed from the following formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (11-4)$$

3. The degrees of freedom for the test are  $n_1 + n_2 - 2$ .

- III. For dependent samples, we assume the population distribution of the paired differences has a mean of 0.
- A. We first compute the mean and the standard deviation of the sample differences.
- B. The value of the test statistic is computed from the following formula:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} \quad (11-5)$$

## PRONUNCIATION KEY

| SYMBOL      | MEANING   | PRONUNCIATION          |
|-------------|---|------------------------|
| $s_p^2$     | Pooled sample variance  | <i>s squared sub p</i> |
| $\bar{x}_1$ | Mean of the first sample  | <i>x bar sub 1</i>     |
| $\bar{x}_2$ | Mean of the second sample   | <i>x bar sub 2</i>     |
| $\bar{d}$   | Mean of the difference between dependent observations               | <i>d bar</i>           |
| $s_d$       | Standard deviation of the difference between dependent observations | <i>s sub d</i>         |

## CHAPTER EXERCISES

17. A recent study focused on the number of times men and women who live alone buy take-out dinner in a month. Assume that the distributions follow the normal probability distribution and the population standard deviations are equal. The information is summarized below.

| Statistic                 | Men   | Women |
|---------------------------|-------|-------|
| Sample mean               | 24.51 | 22.69 |
| Sample standard deviation | 4.48  | 3.86  |
| Sample size               | 35    | 40    |

- At the .01 significance level, is there a difference in the mean number of times men and women order take-out dinners in a month? What is the  $p$ -value?
18. Clark Heter is an industrial engineer at Lyons Products. He would like to determine whether there are more units produced on the night shift than on the day shift. The mean number of units produced by a sample of 54 day-shift workers was 345. The mean number of units produced by a sample of 60 night-shift workers was 351. Assume the population standard deviation of the number of units produced on the day shift is 21 and on the night shift is 28. Using the .05 significance level, is the number of units produced on the night shift larger?
19. Fry Brothers Heating and Air Conditioning Inc. employs Larry Clark and George Murnen to make service calls to repair furnaces and air-conditioning units in homes. Tom Fry, the owner, would like to know whether there is a difference in the mean number of service calls they make per day. A random sample of 40 days last year showed that Larry Clark made an average of 4.77 calls per day. For a sample of 50 days, George Murnen made an average of 5.02 calls per day. Assume the population standard deviation is 1.05 calls per day for Larry Clark and 1.23 calls per day for George Murnen. At the .05 significance level, is there a difference in the mean number of calls per day between the two employees? What is the  $p$ -value?
20. A coffee manufacturer is interested in whether the mean daily consumption of regular-coffee drinkers is less than that of decaffeinated-coffee drinkers. Assume the population standard deviation is 1.20 cups per day for those drinking regular coffee and 1.36 cups per day for those drinking decaffeinated coffee. A random sample of 50 regular-coffee drinkers showed a mean of 4.35 cups per day. A sample of 40 decaffeinated-coffee

drinkers showed a mean of 5.84 cups per day. Use the .01 significance level. Compute the  $p$ -value.

21. A cell phone company offers two plans to its subscribers. At the time new subscribers sign up, they are asked to provide some demographic information. The mean yearly income for a sample of 40 subscribers to Plan A is \$57,000 with a standard deviation of \$9,200. For a sample of 30 subscribers to Plan B, the mean income is \$61,000 with a standard deviation of \$7,100. At the .05 significance level, is it reasonable to conclude the mean income of those selecting Plan B is larger? What is the  $p$ -value?
22. A computer manufacturer offers technical support that is available 24 hours a day, 7 days a week. Timely resolution of these calls is important to the company's image. For 35 calls that were related to software, technicians resolved the issues in a mean time of 18 minutes with a standard deviation of 4.2 minutes. For 45 calls related to hardware, technicians resolved the problems in a mean time of 15.5 minutes with a standard deviation of 3.9 minutes. At the .05 significance level, does it take longer to resolve software issues? What is the  $p$ -value?
23. The owner of Bun 'N' Run Hamburgers wishes to compare the sales per day at two locations. The mean number sold for 10 randomly selected days at the Northside site was 83.55, and the standard deviation was 10.50. For a random sample of 12 days at the Southside location, the mean number sold was 78.80 and the standard deviation was 14.25. At the .05 significance level, is there a difference in the mean number of hamburgers sold at the two locations? What is the  $p$ -value?
24. **FILE** Two of the teams competing in the *America's Cup* race are Team Oracle U.S.A. and Land Rover BAR. They race their boats over a part of the course several times. Below are a sample of times in minutes for each boat. Assume the population standard deviations are the same. At the .05 significance level, can we conclude that there is a difference in their mean times?

| Boat           | Time (minutes) |      |      |      |      |      |      |      |      |      |      |      |
|----------------|----------------|------|------|------|------|------|------|------|------|------|------|------|
| Land Rover BAR | 12.9           | 12.5 | 11.0 | 13.3 | 11.2 | 11.4 | 11.6 | 12.3 | 14.2 | 11.3 |      |      |
| Team Oracle    | 14.1           | 14.1 | 14.2 | 17.4 | 15.8 | 16.7 | 16.1 | 13.3 | 13.4 | 13.6 | 10.8 | 19.0 |

25. **FILE** The manufacturer of an MP3 player wanted to know whether a 10% reduction in price is enough to increase the sales of its product. To investigate, the owner randomly selected eight outlets and sold the MP3 player at the reduced price. At seven randomly selected outlets, the MP3 player was sold at the regular price. Reported below is the number of units sold last month at the regular and reduced prices at the randomly selected outlets. At the .01 significance level, can the manufacturer conclude that the price reduction resulted in an increase in sales?

|                      |     |     |     |     |     |     |     |     |
|----------------------|-----|-----|-----|-----|-----|-----|-----|-----|
| <b>Regular price</b> | 138 | 121 | 88  | 115 | 141 | 125 | 96  |     |
| <b>Reduced price</b> | 128 | 134 | 152 | 135 | 114 | 106 | 112 | 120 |

26. **FILE** A number of minor automobile accidents occur at various high-risk intersections in Teton County, despite traffic lights. The Traffic Department claims that a modification in the type of light will reduce these accidents. The county commissioners have agreed to a proposed experiment. Eight intersections were chosen at random, and the lights at those intersections were modified. The numbers of minor accidents during a six-month period before and after the modifications were:

|                     | Number of Accidents |   |   |   |   |   |   |    |
|---------------------|---------------------|---|---|---|---|---|---|----|
|                     | A                   | B | C | D | E | F | G | H  |
| Before modification | 5                   | 7 | 6 | 4 | 8 | 9 | 8 | 10 |
| After modification  | 3                   | 7 | 7 | 0 | 4 | 6 | 8 | 2  |

At the .01 significance level, is it reasonable to conclude that the modification reduced the number of traffic accidents?



27. **FILE** Lester Hollar is vice president for human resources for a large manufacturing company. In recent years, he has noticed an increase in absenteeism that he thinks is related to the general health of the employees. Four years ago, in an attempt to improve the situation, he began a fitness program in which employees exercise during their lunch hour. To evaluate the program, he selected a random sample of eight participants and found the number of days each was absent in the six months before the exercise program began and in the six months following the exercise program. Below are the results. At the .05 significance level, can he conclude that the number of absences has declined? Estimate the  $p$ -value.

| Employee  | Before | After |
|-----------|--------|-------|
| Bauman    | 6      | 5     |
| Briggs    | 6      | 2     |
| Dottellis | 7      | 1     |
| Lee       | 7      | 3     |
| Perralt   | 4      | 3     |
| Rielly    | 3      | 6     |
| Steinmetz | 5      | 3     |
| Stoltz    | 6      | 7     |

28. **FILE** The president of the American Insurance Institute wants to compare the yearly costs of auto insurance offered by two leading companies. He selects a sample of 15 families, some with only a single insured driver, others with several teenage drivers, and pays each family a stipend to contact the two companies and ask for a price quote. To make the data comparable, certain features, such as the deductible amount and limits of liability, are standardized. The data for the sample of families and their two insurance quotes are reported below. At the .10 significance level, can we conclude that there is a difference in the amounts quoted?

| Family   | Midstates<br>Car Insurance | Gecko<br>Mutual Insurance |
|----------|----------------------------|---------------------------|
| Becker   | \$2,090                    | \$1,610                   |
| Berry    | 1,683                      | 1,247                     |
| Cobb     | 1,402                      | 2,327                     |
| Debuck   | 1,830                      | 1,367                     |
| DuBrul   | 930                        | 1,461                     |
| Eckroate | 697                        | 1,789                     |
| German   | 1,741                      | 1,621                     |
| Glasson  | 1,129                      | 1,914                     |
| King     | 1,018                      | 1,956                     |
| Kucic    | 1,881                      | 1,772                     |
| Meredith | 1,571                      | 1,375                     |
| Obeid    | 874                        | 1,527                     |
| Price    | 1,579                      | 1,767                     |
| Phillips | 1,577                      | 1,636                     |
| Tresize  | 860                        | 1,188                     |

29. Fairfield Homes is developing two parcels near Pigeon Forge, Tennessee. In order to test different advertising approaches, it uses different media to reach potential buyers. The mean annual family income for 15 people making inquiries at the first

development is \$150,000, with a standard deviation of \$40,000. A corresponding sample of 25 people at the second development had a mean of \$180,000, with a standard deviation of \$30,000. Assume the population standard deviations are the same. At the .05 significance level, can Fairfield conclude that the population means are different?

30. A candy company taste-tested two chocolate bars, one with almonds and one without almonds. A panel of testers rated the bars on a scale of 0 to 5, with 5 indicating the highest taste rating. Assume the population standard deviations are equal. At the .05 significance level, do the ratings show a difference between chocolate bars with or without almonds?

| With Almonds | Without Almonds |
|--------------|-----------------|
| 3            | 0               |
| 1            | 4               |
| 2            | 4               |
| 3            | 3               |
| 1            | 4               |
| 1            |                 |
| 2            |                 |

31. **FILE** An investigation of the effectiveness of an antibacterial soap in reducing operating room contamination resulted in the accompanying table. The new soap was tested in a sample of eight operating rooms in the greater Seattle area during the last year. The following table reports the contamination levels before and after the use of the soap for each operating room.

|        | Operating Room |     |     |      |      |     |     |      |
|--------|----------------|-----|-----|------|------|-----|-----|------|
|        | A              | B   | C   | D    | E    | F   | G   | H    |
| Before | 6.6            | 6.5 | 9.0 | 10.3 | 11.2 | 8.1 | 6.3 | 11.6 |
| After  | 6.8            | 2.4 | 7.4 | 8.5  | 8.1  | 6.1 | 3.4 | 2.0  |

At the .05 significance level, can we conclude the contamination measurements are lower after use of the new soap?

32. **FILE** The following data on annual rates of return were collected from 11 randomly selected stocks listed on the New York Stock Exchange (“the big board”) and 12 randomly selected stocks listed on NASDAQ. Assume the population standard deviations are the same. At the .10 significance level, can we conclude that the annual rates of return are higher on the big board?

| NYSE | NASDAQ |
|------|--------|
| 15.0 | 8.8    |
| 10.7 | 6.0    |
| 20.2 | 14.4   |
| 18.6 | 19.1   |
| 19.1 | 17.6   |
| 8.7  | 17.8   |
| 17.8 | 15.9   |
| 13.8 | 17.9   |
| 22.7 | 21.6   |
| 14.0 | 6.0    |
| 26.1 | 11.9   |
|      | 23.4   |

33. **FILE** The city of Laguna Beach operates two public parking lots. The Ocean Drive parking lot can accommodate up to 125 cars, and the Rio Rancho parking lot can accommodate up to 130 cars. City planners are considering increasing the size of the lots and changing the fee structure. To begin, the Planning Office would like some information on the number of cars in the lots at various times of the day. A junior planner officer is assigned the task of visiting the two lots at random times of the day and evening and counting the number of cars in the lots. The study lasted over a period of one month. Below is the number of cars in the lots for 25 visits of the Ocean Drive lot and 28 visits of the Rio Rancho lot. Assume the population standard deviations are equal.

| Ocean Drive |     |    |     |     |     |     |    |     |     |     |    |     |
|-------------|-----|----|-----|-----|-----|-----|----|-----|-----|-----|----|-----|
| 89          | 115 | 93 | 79  | 113 | 77  | 51  | 75 | 118 | 105 | 106 | 91 | 54  |
| 63          | 121 | 53 | 81  | 115 | 67  | 53  | 69 | 95  | 121 | 88  | 64 |     |
| Rio Rancho  |     |    |     |     |     |     |    |     |     |     |    |     |
| 128         | 110 | 81 | 126 | 82  | 114 | 93  | 40 | 94  | 45  | 84  | 71 | 74  |
| 92          | 66  | 69 | 100 | 114 | 113 | 107 | 62 | 77  | 80  | 107 | 90 | 129 |
| 105         | 124 |    |     |     |     |     |    |     |     |     |    |     |

Is it reasonable to conclude that there is a difference in the mean number of cars in the two lots? Use the .05 significance level.

34. **FILE** The amount of income spent on housing is an important component of the cost of living. The total costs of housing for homeowners might include mortgage payments, property taxes, and utility costs (water, heat, electricity). An economist selected a sample of 20 homeowners in New England and then calculated these total housing costs as a percent of monthly income, 5 years ago and now. The information is reported below. Is it reasonable to conclude the percent is less now than 5 years ago?

| Homeowner | Five Years Ago | Now | Homeowner  | Five Years Ago | Now |
|-----------|----------------|-----|------------|----------------|-----|
| Holt      | 17%            | 10% | Lozier     | 35%            | 32% |
| Pierse    | 20             | 39  | Cieslinski | 16             | 32  |
| Merenick  | 29             | 37  | Rowatti    | 23             | 21  |
| Lanoué    | 43             | 27  | Koppel     | 33             | 12  |
| Fagan     | 36             | 12  | Rumsey     | 44             | 40  |
| Bobko     | 43             | 41  | McGinnis   | 44             | 42  |
| Kippert   | 45             | 24  | Pierce     | 28             | 22  |
| San Roman | 19             | 26  | Roll       | 29             | 19  |
| Kurimsky  | 49             | 28  | Lang       | 39             | 35  |
| Davison   | 49             | 26  | Miller     | 22             | 12  |

35. **FILE** The CVS Pharmacy located on US 17 in Murrells Inlet has been one of the busiest pharmaceutical retail stores in South Carolina for many years. To try and capture more business in the area, CVS top management opened another store about 6 miles west on SC 707. After a few months, CVS management decided to compare the business volume at the two stores. One way to measure business volume is to count the number of cars in the store parking lots on random days and times. The results of the survey from the last 3 months of the year are reported below. To explain, the first observation was on October 2 at 20:52 military time (8:52 p.m.). At that time there were four cars in the US 17 lot and nine cars in the SC 707 lot. At the .05 significance level, is it reasonable to

conclude that, based on vehicle counts, the US 17 store has more business volume than the SC 707 store?

| Date   | Time  | Vehicle Count |        |
|--------|-------|---------------|--------|
|        |       | US 17         | SC 707 |
| Oct 2  | 20:52 | 4             | 9      |
| Oct 11 | 19:30 | 5             | 7      |
| Oct 15 | 22:08 | 9             | 12     |
| Oct 19 | 11:42 | 4             | 5      |
| Oct 25 | 15:32 | 10            | 8      |
| Oct 26 | 11:02 | 9             | 15     |
| Nov 3  | 11:22 | 13            | 7      |
| Nov 5  | 19:09 | 20            | 3      |
| Nov 8  | 15:10 | 15            | 14     |
| Nov 9  | 13:18 | 15            | 11     |
| Nov 15 | 22:38 | 13            | 11     |
| Nov 17 | 18:46 | 16            | 12     |
| Nov 21 | 15:44 | 17            | 8      |
| Nov 22 | 15:34 | 15            | 3      |
| Nov 27 | 21:42 | 20            | 6      |
| Nov 29 | 9:57  | 17            | 13     |
| Nov 30 | 17:58 | 5             | 9      |
| Dec 3  | 19:54 | 7             | 13     |
| Dec 15 | 18:20 | 11            | 6      |
| Dec 16 | 18:25 | 14            | 15     |
| Dec 17 | 11:08 | 8             | 8      |
| Dec 22 | 21:20 | 10            | 3      |
| Dec 24 | 15:21 | 4             | 6      |
| Dec 25 | 20:21 | 7             | 9      |
| Dec 30 | 14:25 | 19            | 4      |

36. **FILE** A goal of financial literacy for children is to learn how to manage money wisely. One question is: How much money do children have to manage? A recent study by Schnur Educational Research Associates randomly sampled 15 children between 8 and 10 years old and 18 children between 11 and 14 years old and recorded their monthly allowance. Is it reasonable to conclude that the mean allowance received by children between 11 and 14 years is more than the allowance received by children between 8 and 10 years? Use the .01 significance level. What is the  $p$ -value?

| 8–10 Years | 11–14 Years | 8–10 Years | 11–14 Years |
|------------|-------------|------------|-------------|
| 26         | 49          | 26         | 41          |
| 33         | 44          | 25         | 38          |
| 30         | 42          | 27         | 44          |
| 26         | 38          | 29         | 39          |
| 34         | 39          | 34         | 50          |
| 26         | 41          | 32         | 49          |
| 27         | 39          |            | 41          |
| 27         | 38          |            | 42          |
| 30         | 38          |            | 30          |

## DATA ANALYTICS

(The data for these exercises are available in Connect.)

37. **FILE** The North Valley Real Estate data report information on the homes sold last year.
- At the .05 significance level, can we conclude that there is a difference in the mean selling price of homes with a pool and homes without a pool?
  - At the .05 significance level, can we conclude that there is a difference in the mean selling price of homes with an attached garage and homes without an attached garage?
  - At the .05 significance level, can we conclude that there is a difference in the mean selling price of homes that are in default on the mortgage?
38. **FILE** Refer to the Baseball 2016 data, which report information on the 30 Major League Baseball teams for the 2016 season.
- At the .05 significance level, can we conclude that there is a difference in the mean salary of teams in the American League versus teams in the National League?
  - At the .05 significance level, can we conclude that there is a difference in the mean home attendance of teams in the American League versus teams in the National League?
  - Compute the mean and the standard deviation of the number of wins for the 10 teams with the highest salaries. Do the same for the 10 teams with the lowest salaries. At the .05 significance level, is there a difference in the mean number of wins for the two groups? At the .05 significance level, is there a difference in the mean attendance for the two groups?
39. **FILE** Refer to the Lincolnville School District bus data. Is there a difference in the mean maintenance cost for the diesel versus the gasoline buses? Use the .05 significance level.

## PRACTICE TEST

### Part 1—Objective

- The hypothesized *difference* between two population means is \_\_\_\_\_. (0, not equal to 1, 1, at least 1)
- If the population standard deviations are known for a test of differences of two independent population means, the test statistic is a \_\_\_\_\_ statistic.
- When sampled items from two populations are classified as “success” or “failure,” the hypothesis test is for differences in population \_\_\_\_\_.
- For two independent populations, sample standard deviations are pooled to compute a single estimate of the \_\_\_\_\_. (population mean, population standard deviation, population proportion, z value)
- A hypothesis test of differences between two *dependent* populations is based on a single population of mean \_\_\_\_\_. (differences, populations, standard deviations, *t* values)
- The test statistic for a hypothesis test of differences between two dependent populations follows the \_\_\_\_\_ distribution.
- Degrees of freedom for a hypothesis test of differences between two dependent populations is \_\_\_\_\_.
- For dependent samples, observations are matched or \_\_\_\_\_.
- For independent samples, the two samples are different or \_\_\_\_\_.
- In a statistics class, for each student the percentage correct on Exam 1 is subtracted from the percentage correct on Exam 2. This is an example of \_\_\_\_\_ samples.

**Part 2—Problems**

For each of these problems, use the six-step hypothesis-testing procedure.

1. The city of Myrtle Beach is comparing two taxi companies to see whether they differ in the mean miles traveled per week. The data are summarized in the following table. Using the .05 significance level, is there a difference in the mean miles traveled?

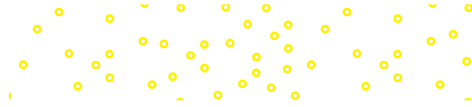
|                    | Yellow Cab | Horse and Buggy Cab |
|--------------------|------------|---------------------|
| Mean miles         | 837        | 797                 |
| Standard deviation | 30         | 40                  |
| Sample size        | 14         | 12                  |

2. Dial Soap Company developed a new soap for men and test-marketed the product in two cities. The sample information is reported below. At the .05 significance level, can we conclude there is a difference in the proportion that liked the new soap in the two cities?

| City       | Liked New Soap | Number Sampled |
|------------|----------------|----------------|
| Erie, PA   | 128            | 300            |
| Tustin, CA | 149            | 400            |

# 12

# Analysis of Variance



©Denys Prykhodov/Shutterstock

- ▲ **ONE VARIABLE THAT GOOGLE** uses to rank pages on the Internet is page speed, the time it takes for a web page to load into your browser. A source for women's clothing is redesigning their page to improve the images that show its products and to reduce its load time. The new page is clearly faster, but initial tests indicate there is more variation in the time to load. A sample of 16 different load times showed that the standard deviation of the load time was 22 hundredths of a second for the new page and 12 hundredths of a second for the current page. At the .05 significance level, can we conclude that there is more variation in the load time of the new page? (See Exercise 16 and **LO12-1**.)

## LEARNING OBJECTIVES

---

*When you have completed this chapter, you will be able to:*

- LO12-1** Apply the  $F$  distribution to test a hypothesis that two population variances are equal.
- LO12-2** Use ANOVA to test a hypothesis that three or more population means are equal.
- LO12-3** Use confidence intervals to test and interpret differences between pairs of population means.



## INTRODUCTION

In this chapter, we continue our discussion of hypothesis testing. Recall that in Chapters 10 and 11 we examined the idea of hypothesis testing. We described the case where a sample was selected from the population. We used the  $z$  distribution (the standard normal distribution) or the  $t$  distribution to determine whether it was reasonable to conclude that the population mean was equal to a specified value. We tested whether two population means are the same. In this chapter, we expand our idea of hypothesis tests. We describe a test for variances and then a test that simultaneously compares several population means to determine if they are equal.

### LO12-1

Apply the  $F$  distribution to test a hypothesis that two population variances are equal.

## COMPARING TWO POPULATION VARIANCES

In Chapter 11, we tested hypotheses about equal population means. The tests differed based on our assumptions regarding whether the population standard deviations or variances were equal or unequal. In this chapter, the assumption about equal population variances is also important. Here we describe a method to statistically test this assumption. The test is based on the  $F$  distribution.

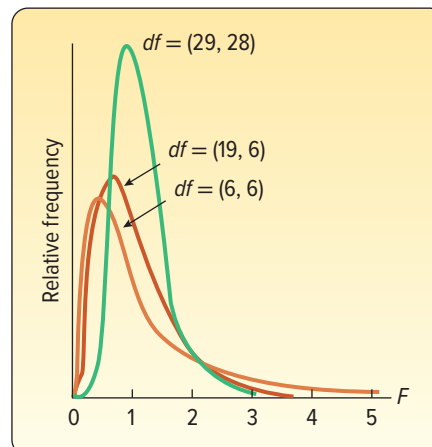
### The $F$ Distribution

The probability distribution used in this chapter is the  $F$  distribution. It was named to honor Sir Ronald Fisher, one of the founders of modern-day statistics. The test statistic for several situations follows this probability distribution. It is used to test whether two samples are from populations having equal variances, and it is also applied when we want to compare several population means simultaneously. The simultaneous comparison of several population means is called **analysis of variance (ANOVA)**. In both of these situations, the populations must follow a normal distribution, and the data must be at least interval-scale.

**ANALYSIS OF VARIANCE (ANOVA)** A technique used to test simultaneously whether the means of several populations are equal. It uses the  $F$  distribution as the distribution of the test statistic.

What are the characteristics of the  $F$  distribution?

1. **There is a family of  $F$  distributions.** A particular member of the family is determined by two parameters: the degrees of freedom in the numerator and the degrees of freedom in the denominator. The shape of the distribution is illustrated by the following graph. There is one  $F$  distribution for the combination of 29 degrees of freedom in the numerator ( $df$ ) and 28 degrees of freedom in the denominator. There is another  $F$  distribution for 19 degrees of freedom in the numerator and 6 degrees of freedom in the denominator. The final distribution shown has 6 degrees of freedom





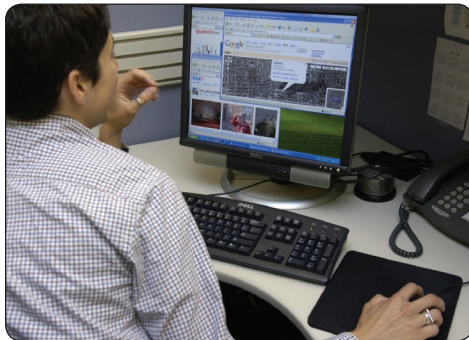
in the numerator and 6 degrees of freedom in the denominator. We will describe the concept of degrees of freedom later in the chapter. Note that the shapes of the distributions change as the degrees of freedom change.

2. **The  $F$  distribution is continuous.** This means that the value of  $F$  can assume an infinite number of values between zero and positive infinity.
3. **The  $F$ -statistic cannot be negative.** The smallest value  $F$  can assume is 0.
4. **The  $F$  distribution is positively skewed.** The long tail of the distribution is to the right-hand side. As the number of degrees of freedom increases in both the numerator and denominator, the distribution approaches a normal distribution.
5. **The  $F$  distribution is asymptotic.** As the values of  $F$  increase, the distribution approaches the horizontal axis but never touches it. This is similar to the behavior of the normal probability distribution, described in Chapter 7.

## Testing a Hypothesis of Equal Population Variances

The first application of the  $F$  distribution that we describe occurs when we test the hypothesis that the variance of one normal population equals the variance of another normal population. The following examples will show the use of the test:

- A health services corporation manages two hospitals in Knoxville, Tennessee: St. Mary's North and St. Mary's South. In each hospital, the mean waiting time in the Emergency Department is 42 minutes. The hospital administrator believes that the St. Mary's North Emergency Department has more variation in waiting time than St. Mary's South.
  - The mean rate of return on two types of common stock may be the same, but there may be more variation in the rate of return in one than the other. A sample of 10 technology and 10 utility stocks shows the same mean rate of return, but there is likely more variation in the technology stocks.
  - An online newspaper found that men and women spend about the same amount of time per day accessing news apps. However, the same report indicated the times of men had nearly twice as much variation compared to the times of women.



©McGraw-Hill Education/John Flournoy, photographer

The  $F$  distribution is also used to test the assumption that the variances of two normal populations are equal. Recall that in the previous chapter the  $t$  test to investigate whether the means of two independent populations differed assumes that the variances of the two normal populations are the same. See this list of assumptions on page 313. The  $F$  distribution is used to test the assumption that the variances are equal.

To compare two population variances, we first state the null hypothesis. The null hypothesis is that the variance of one normal population,  $\sigma_1^2$ , equals the variance of another normal population,  $\sigma_2^2$ . The alternate hypothesis is that the variances differ. In this instance, the null hypothesis and the alternate hypothesis are:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

To conduct the test, we select a random sample of observations,  $n_1$ , from one population and a random sample of observations,  $n_2$ , from the second population. The test statistic is defined as follows.

**TEST STATISTIC FOR COMPARING TWO VARIANCES**

$$F = \frac{s_1^2}{s_2^2}$$

**(12-1)**

The terms  $s_1^2$  and  $s_2^2$  are the respective sample variances. If the null hypothesis is true, the test statistic follows the  $F$  distribution with  $n_1 - 1$  and  $n_2 - 1$  degrees of freedom. To reduce the size of the table of critical values, the *larger* sample variance is placed in the numerator; hence, the tabled  $F$  ratio is always larger than 1.00. Thus, the right-tail critical value is the only one required. The critical value of  $F$  for a two-tailed test is found by dividing the significance level in half ( $\alpha/2$ ) and then referring to the appropriate degrees of freedom in Appendix B.6. An example will illustrate.

### EXAMPLE

Lammers Limos offers limousine service from Government Center in downtown Toledo, Ohio, to Metro Airport in Detroit. Sean Lammers, president of the company, is considering two routes. One is via U.S. 25 and the other via I-75. He wants to study the time it takes to drive to the airport using each route and then compare the results. He collected the following sample data, which is reported in minutes. Using the .10 significance level, is there a difference in the variation in the driving times for the two routes?



©egd/Shutterstock

| U.S. Route 25 | Interstate 75 |
|---------------|---------------|
| 52            | 59            |
| 67            | 60            |
| 56            | 61            |
| 45            | 51            |
| 70            | 56            |
| 54            | 63            |
| 64            | 57            |
|               | 65            |

### SOLUTION

The mean driving times along the two routes are nearly the same. The mean time is 58.29 minutes for the U.S. 25 route and 59.0 minutes along the I-75 route. However, in evaluating travel times, Mr. Lammers is also concerned about the variation in the travel times. The first step is to compute the two sample variances. We'll use formula (3–8) to compute the sample standard deviations. To obtain the sample variances, we square the standard deviations.

#### U.S. ROUTE 25

$$\bar{x} = \frac{\sum x}{n} = \frac{408}{7} = 58.29 \quad s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{485.43}{7 - 1}} = 8.9947$$

#### INTERSTATE 75

$$\bar{x} = \frac{\sum x}{n} = \frac{472}{8} = 59.00 \quad s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{134}{8 - 1}} = 4.3753$$

There is more variation, as measured by the standard deviation, in the U.S. 25 route than in the I-75 route. This is consistent with his knowledge of the two routes; the U.S. 25 route contains more stoplights, whereas I-75 is a limited-access interstate highway. However, the I-75 route is several miles longer. It is important that the service

offered be both timely and consistent, so he decides to conduct a statistical test to determine whether there really is a difference in the variation of the two routes.

We use the six-step hypothesis test procedure.

**Step 1:** We begin by stating the null hypothesis and the alternate hypothesis. The test is two-tailed because we are looking for a difference in the variation of the two routes. We are *not* trying to show that one route has more variation than the other. For this example/solution, the subscript 1 indicates information for U.S. 25; the subscript 2 indicates information for I-75.

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

**Step 2:** We selected the .10 significance level.

**Step 3:** The appropriate test statistic follows the  $F$  distribution.

**Step 4:** The critical value is obtained from Appendix B.6, a portion of which is reproduced as Table 12–1. Because we are conducting a two-tailed test, the tabled significance level is .05, found by  $\alpha/2 = .10/2 = .05$ . There are  $n_1 - 1 = 7 - 1 = 6$  degrees of freedom in the numerator and  $n_2 - 1 = 8 - 1 = 7$  degrees of freedom in the denominator. To find the critical value, move horizontally across the top portion of the  $F$  table (Table 12–1 or Appendix B.6) for the .05 significance level to 6 degrees of freedom in the numerator. Then move down that column to the critical value opposite 7 degrees of freedom in the denominator. The critical value is 3.87. Thus, the decision rule is: Reject the null hypothesis if the ratio of the sample variances exceeds 3.87.

**TABLE 12–1** Critical Values of the  $F$  Distribution,  $\alpha = .05$

| Degrees of Freedom for Denominator | Degrees of Freedom for Numerator |      |      |      |
|------------------------------------|----------------------------------|------|------|------|
|                                    | 5                                | 6    | 7    | 8    |
| 1                                  | 230                              | 234  | 237  | 239  |
| 2                                  | 19.3                             | 19.3 | 19.4 | 19.4 |
| 3                                  | 9.01                             | 8.94 | 8.89 | 8.85 |
| 4                                  | 6.26                             | 6.16 | 6.09 | 6.04 |
| 5                                  | 5.05                             | 4.95 | 4.88 | 4.82 |
| 6                                  | 4.39                             | 4.28 | 4.21 | 4.15 |
| 7                                  | 3.97                             | 3.87 | 3.79 | 3.73 |
| 8                                  | 3.69                             | 3.58 | 3.50 | 3.44 |
| 9                                  | 3.48                             | 3.37 | 3.29 | 3.23 |
| 10                                 | 3.33                             | 3.22 | 3.14 | 3.07 |

**Step 5:** Next we compute the ratio of the two sample variances, determine the value of the test statistic, and make a decision regarding the null hypothesis. Note that formula (12–1) refers to the sample *variances*, but we calculated the sample *standard deviations*. We need to square the standard deviations to determine the variances.

$$F = \frac{s_1^2}{s_2^2} = \frac{(8.9947)^2}{(4.3753)^2} = 4.23$$

The decision is to reject the null hypothesis because the computed  $F$ -value (4.23) is larger than the critical value (3.87).

**Step 6:** We conclude there is a difference in the variation in the time to travel the two routes. Mr. Lammers will want to consider this in his scheduling.

The usual practice is to determine the  $F$  ratio by putting the larger of the two sample variances in the numerator. This will force the  $F$  ratio to be at least 1.00. This allows us to always use the right tail of the  $F$  distribution, thus avoiding the need for more extensive  $F$  tables.

A logical question arises: Is it possible to conduct one-tailed tests? For example, suppose in the previous example we suspected that the variance of the times using the U.S. 25 route,  $\sigma_1^2$ , is larger than the variance of the times along the I-75 route,  $\sigma_2^2$ . We would state the null and the alternate hypotheses as

$$H_0: \sigma_1^2 \leq \sigma_2^2$$

$$H_1: \sigma_1^2 > \sigma_2^2$$

The test statistic is computed as  $s_1^2/s_2^2$ . Notice that we labeled the population with the suspected large variance as population 1. So  $s_1^2$  appears in the numerator. The  $F$  ratio will be larger than 1.00, so we can use the upper tail of the  $F$  distribution. Under these conditions, it is not necessary to divide the significance level in half. Because Appendix B.6 gives us only the .05 and .01 significance levels, we are restricted to these levels for one-tailed tests and .10 and .02 for two-tailed tests unless we consult a more complete table or use statistical software to compute the  $F$ -statistic.

The Excel software has a procedure to perform a test of variances. Below is the output. The computed value of  $F$  is the same as that determined by using formula (12–1). The result of the one-tail hypothesis test is to reject the null hypothesis. The  $F$  of 4.23 is greater than the critical value of 3.87. Also, the  $p$ -value is less than 0.05. We conclude the variance of travel times on U.S. 25 is greater than the variance of travel times on I-75.

| Variance Test |         |               |   |                                 |         |               |   |
|---------------|---------|---------------|---|---------------------------------|---------|---------------|---|
|               | A       | B             | C | D                               | E       | F             | G |
| 1             | U.S. 25 | Interstate 75 |   | F-Test Two-Sample for Variances |         |               |   |
| 2             | 52      | 59            |   |                                 | U.S. 25 | Interstate 75 |   |
| 3             | 67      | 60            |   | Mean                            | 58.29   | 59.00         |   |
| 4             | 56      | 61            |   | Variance                        | 80.90   | 19.14         |   |
| 5             | 45      | 51            |   | Observations                    | 7.00    | 8.00          |   |
| 6             | 70      | 56            |   | df                              | 6.00    | 7.00          |   |
| 7             | 54      | 63            |   | F                               | 4.23    |               |   |
| 8             | 64      | 57            |   | P(F<=f) one-tail                | 0.04    |               |   |
| 9             |         | 65            |   | F Critical one-tail             | 3.87    |               |   |
| 10            |         |               |   |                                 |         |               |   |

Source: Microsoft Excel

## SELF-REVIEW 12-1



Steele Electric Products Inc. assembles cell phones. For the last 10 days, Mark Nagy completed a mean of 39 phones per day, with a standard deviation of 2 per day. Debbie Richmond completed a mean of 38.5 phones per day, with a standard deviation of 1.5 per day. At the .05 significance level, can we conclude that there is more variation in Mark's daily production?

## EXERCISES

1. What is the critical  $F$ -value when the sample size for the numerator is six and the sample size for the denominator is four? Use a two-tailed test and the .10 significance level.
2. What is the critical  $F$ -value when the sample size for the numerator is four and the sample size for the denominator is seven? Use a one-tailed test and the .01 significance level.
3. The following hypotheses are given.

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

A random sample of eight observations from the first population resulted in a standard deviation of 10. A random sample of six observations from the second population resulted in a standard deviation of 7. At the .02 significance level, is there a difference in the variation of the two populations?

4. The following hypotheses are given.

$$H_0: \sigma_1^2 \leq \sigma_2^2$$

$$H_1: \sigma_1^2 > \sigma_2^2$$

A random sample of five observations from the first population resulted in a standard deviation of 12. A random sample of seven observations from the second population showed a standard deviation of 7. At the .01 significance level, is there more variation in the first population?

5. Arbitron Media Research Inc. conducted a study of the iPod listening habits of men and women. One facet of the study involved the mean listening time. It was discovered that the mean listening time for a sample of 10 men was 35 minutes per day. The standard deviation was 10 minutes per day. The mean listening time for a sample of 12 women was also 35 minutes, but the standard deviation of the sample was 12 minutes. At the .10 significance level, can we conclude that there is a difference in the variation in the listening times for men and women?
6. A stockbroker at Critical Securities reported that the mean rate of return on a sample of 10 oil stocks was 12.6% with a standard deviation of 3.9%. The mean rate of return on a sample of 8 utility stocks was 10.9% with a standard deviation of 3.5%. At the .05 significance level, can we conclude that there is more variation in the oil stocks?

### LO12-2

Use ANOVA to test a hypothesis that three or more population means are equal.

## ANOVA: ANALYSIS OF VARIANCE

The  $F$  distribution is used to perform a wide variety of hypothesis tests. For example, when testing the equality of three or more population means, the analysis of variance (ANOVA) technique is used and the  $F$ -statistic is used as the test statistic.

### ANOVA Assumptions

Using ANOVA to test the equality of three or more population means requires that three assumptions are true:

1. The populations follow the normal distribution.
2. The populations have equal standard deviations ( $\sigma$ ).
3. The populations are independent.

When these conditions are met,  $F$  is used as the distribution of the test statistic.

Why do we need to study ANOVA? Why can't we just use the test of differences in population means discussed in the previous chapter? We could compare the population means two at a time. The major reason is the unsatisfactory buildup of Type I error. To explain further, suppose we have four different methods (A, B, C, and D) of training new recruits to be firefighters. We randomly assign each of the 40 recruits in this year's class to one of the four methods. At the end of the training program, we administer a test to measure understanding of firefighting techniques to the four groups. The question is: Is there a difference in the mean test scores among the four groups? An answer to this question will allow us to compare the four training methods.

Using the  $t$  distribution to compare the four population means, we would have to conduct six different  $t$  tests. That is, we would need to compare the mean scores for the four methods as follows: A versus B, A versus C, A versus D, B versus C, B versus D, and C versus D. For each  $t$  test, suppose we choose an  $\alpha = .05$ . Therefore, the probability of

a Type I error, rejecting the null when it is true, is .05. The complement is the probability of .95 that we do not reject the null when it is true. Because we conduct six separate (independent) tests, the probability that all six tests result in correct decisions is:

$$P(\text{All correct}) = (.95)(.95)(.95)(.95)(.95)(.95) = .735$$

To find the probability of at least one error due to sampling, we subtract this result from 1. Thus, the probability of at least one incorrect decision due to sampling is  $1 - .735 = .265$ . To summarize, if we conduct six independent tests using the  $t$  distribution, the likelihood of rejecting a true null hypothesis because of sampling error is an unsatisfactory .265. The ANOVA technique allows us to compare population means simultaneously at a selected significance level. It avoids the buildup of Type I error associated with testing many hypotheses.

ANOVA was first developed for applications in agriculture, and many of the terms related to that context remain. In particular, the term *treatment* is used to identify the different populations being examined. For example, treatment refers to how a plot of ground was treated with a particular type of fertilizer. The following illustration will clarify the term *treatment* and demonstrate an application of ANOVA.

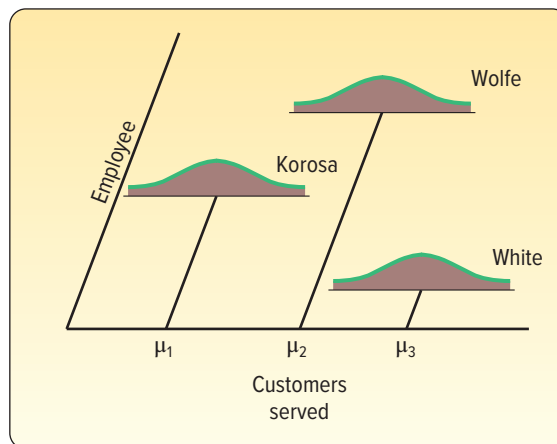
### EXAMPLE

Joyce Kuhlman manages a regional financial center. She wishes to compare the productivity, as measured by the number of customers served, among three employees. Four days are randomly selected, and the number of customers served by each employee is recorded. The results are:

| Wolfe | White | Korosa |
|-------|-------|--------|
| 55    | 66    | 47     |
| 54    | 76    | 51     |
| 59    | 67    | 46     |
| 56    | 71    | 48     |

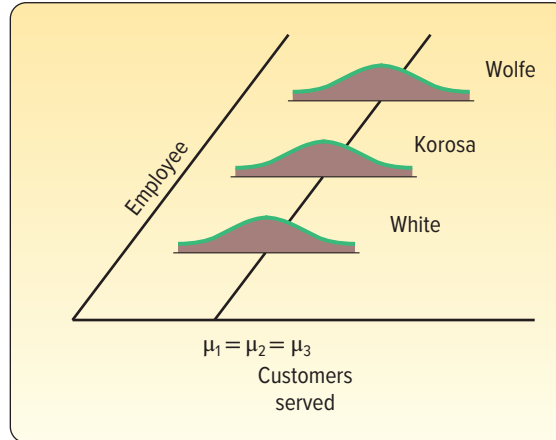
### SOLUTION

Is there a difference in the mean number of customers served? Chart 12–1 illustrates how the populations would appear if there were a difference in the treatment means. Note that the populations follow the normal distribution and the variation in each population is the same. However, the means are *not* the same.



**CHART 12–1** Case Where Treatment Means Are Different

Suppose there is no difference in the treatment means. This would indicate that the population means are the same. This is shown in Chart 12–2. Note again that the populations follow the normal distribution and the variation in each of the populations is the same.



**CHART 12–2** Case Where Treatment Means Are the Same

## The ANOVA Test

How does the ANOVA test work? Recall that we want to determine whether the various sample means came from a single population or populations with different means. We actually compare these sample means through their variances. To explain, on page 340 we listed the assumptions required for ANOVA. One of those assumptions was that the standard deviations of the various normal populations had to be the same. We take advantage of this requirement in the ANOVA test. The underlying strategy is to estimate the population variance (standard deviation squared) two ways and then find the ratio of these two estimates. If this ratio is about 1, then logically the two estimates are the same, and we conclude that the population means are the same. If the ratio is quite different from 1, then we conclude that the population means are not the same. The  $F$  distribution serves as a referee by indicating when the ratio of the sample variances is too much greater than 1 to have occurred by chance.

Refer to the example/solution in the previous section. The manager wants to determine whether there is a difference in the mean number of customers served. To begin, find the overall mean of the 12 observations. It is 58, found by  $(55 + 54 + \dots + 48)/12$ . Next, for each of the 12 observations find the difference between the particular value and the overall mean. Each of these differences is squared and these squares summed. This term is called the **total variation**.

**TOTAL VARIATION** The sum of the squared differences between each observation and the overall mean.

In our example, the total variation is 1,082, found by  $(55 - 58)^2 + (54 - 58)^2 + \dots + (48 - 58)^2$ .

Next, break this total variation into two components: variation due to the **treatment variation** and **random variation**.

**TREATMENT VARIATION** The sum of the squared differences between each treatment mean and the grand or overall mean.

The variation due to treatments is also called variation between treatment means. In this example, we first square the difference between each treatment mean and the overall mean. The mean for Wolfe is 56 customers, found by  $(55 + 54 + 59 + 56)/4$ . The other means are 70 and 48, respectively. Then, each of the squared differences is multiplied by the number of observations in each treatment. In this case, the value is 4. Last, these values are summed together. This term is 992. The sum of the squares due to the treatments is:

$$4(56 - 58)^2 + 4(70 - 58)^2 + 4(48 - 58)^2 = 992$$

If there is considerable variation among the treatment means compared to the overall mean, it is logical that this term will be a large value. If the treatment means are similar, this value will be small. The smallest possible value would be zero. This would occur when all the treatment means are the same. In this case, all the treatment means would also equal the overall mean.

The other source of variation is referred to as **random variation**, or the error component.

**RANDOM VARIATION** The sum of the squared differences between each observation and its treatment mean.

In the example, this term is the sum of the squared differences between each value and the mean for each treatment or employee. This is also called the variation within the treatments. The error variation is 90.

$$(55 - 56)^2 + (54 - 56)^2 + \dots + (48 - 48)^2 = 90$$

We determine the test statistic, which is the ratio of the two estimates of the population variance, from the following equation.

$$F = \frac{\text{Estimate of the population variance based on the differences between the treatment means}}{\text{Estimate of the population variance based on the variation within the treatments}}$$

Our first estimate of the population variance is based on the treatments, that is, the difference *between* the means. It is  $992/2$ . Why did we divide by 2? Recall from Chapter 3, to find a sample variance [see formula (3-7)], we divide by the number of observations minus one. In this case, there are three treatments, so we divide by 2. Our first estimate of the population variance is  $992/2$ .

The variance estimate *within* the treatments is the random variation divided by the total number of observations less the number of treatments—that is,  $90/(12 - 3)$ . Hence, our second estimate of the population variance is  $90/9$ .

The last step is to take the ratio of these two estimates.

$$F = \frac{992/2}{90/9} = 49.6$$

Because this ratio is quite different from 1, we can conclude that the treatment means are not the same. There is a difference in the mean number of customers served by the three employees.

Here's another example, which deals with samples of different sizes.



### EXAMPLE

The airline industry monitors and tracks customer satisfaction. For example, the industry collects information on overbooking and baggage handling. A group of four carriers hired Brunner Marketing Research Inc. to survey passengers regarding their level of satisfaction with a recent flight. Twenty-five questions offered a range of possible answers: excellent, good, fair, or poor. A response of excellent was given a score of 4, good a 3, fair a 2, and poor a 1. These responses were then totaled, so the total score was an indication of the satisfaction with the flight. The greater the score, the higher the level of satisfaction with the service. The highest possible score was 100.

Brunner randomly selected and surveyed passengers from the four airlines. Below is the sample information. Is there a difference in the mean satisfaction level among the four airlines? Use the .01 significance level.

| Northern | WTA | Pocono | Branson |
|----------|-----|--------|---------|
| 94       | 75  | 70     | 68      |
| 90       | 68  | 73     | 70      |
| 85       | 77  | 76     | 72      |
| 80       | 83  | 78     | 65      |
|          | 88  | 80     | 74      |
|          |     | 68     | 65      |
|          |     | 65     |         |

### SOLUTION

We will use the six-step hypothesis-testing procedure.

**Step 1: State the null hypothesis and the alternate hypothesis.** The null hypothesis is that the mean scores are the same for the four airlines.

$$H_0: \mu_N = \mu_W = \mu_P = \mu_B$$

The alternate hypothesis is that the mean scores are not all the same for the four airlines.

$$H_1: \text{The mean scores are not all equal.}$$

We can also think of the alternate hypothesis as “at least two mean scores are not equal.”

If the null hypothesis is not rejected, we conclude that there is no difference in the mean scores for the four airlines. If  $H_0$  is rejected, we conclude that there is a difference in at least one pair of mean scores, but at this point we do not know which pair or how many pairs differ.

**Step 2: Select the level of significance.** We selected the .01 significance level.

**Step 3: Determine the test statistic.** The test statistic follows the  $F$  distribution.

**Step 4: Formulate the decision rule.** To determine the decision rule, we need the critical value. The critical value for the  $F$ -statistic is found in Appendix B.6. The critical values for the .05 significance level are found in Appendix B.6A and the .01 significance level in Appendix B.6B. To use this table, we need to know the degrees of freedom in the numerator and the denominator. The degrees of freedom in the numerator equal the number of treatments, designated as  $k$ , minus 1.

The degrees of freedom in the denominator are the total number of observations,  $n$ , minus the number of treatments. For this problem, there are four treatments and a total of 22 observations.

$$\text{Degrees of freedom in the numerator} = k - 1 = 4 - 1 = 3$$

$$\text{Degrees of freedom in the denominator} = n - k = 22 - 4 = 18$$

Refer to the critical values of  $F$  for the .01 significance level in Appendix B.6B. Move horizontally across the top of the page to 3 degrees of freedom in the numerator. Then move down that column to the row with 18 degrees of freedom. The value at this intersection is 5.09. So the decision rule is to reject  $H_0$  if the computed value of  $F$  exceeds 5.09.

**Step 5: Select the sample, perform the calculations, and make a decision.** It is convenient to summarize the calculations of the  $F$ -statistic in an **ANOVA table**. The format for an ANOVA table is as follows. Statistical software packages also use this format.

| ANOVA Table         |                |                    |                     |           |
|---------------------|----------------|--------------------|---------------------|-----------|
| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square         | $F$       |
| Treatments          | SST            | $k - 1$            | $SST/(k - 1) = MST$ | $MST/MSE$ |
| Error               | SSE            | $n - k$            | $SSE/(n - k) = MSE$ |           |
| Total               | SS total       | $n - 1$            |                     |           |

There are three values, or sum of squares, used to compute the test statistic  $F$ . You can determine these values by obtaining SS total and SSE, then finding SST by subtraction. The SS total term is the total variation, SST is the variation due to the treatments, and SSE is the variation within the treatments or the random error.

We usually start the process by finding SS total. This is the sum of the squared differences between each observation and the overall mean. The formula for finding SS total is:

$$SS \text{ total} = \sum(x - \bar{x}_G)^2 \quad (12-2)$$

where:

$x$  is each sample observation.

$\bar{x}_G$  is the overall or grand mean.

Next determine SSE or the sum of the squared errors. This is the sum of the squared differences between each observation and its respective treatment mean. The formula for finding SSE is:

$$SSE = \sum(x - \bar{x}_c)^2 \quad (12-3)$$

where:

$\bar{x}_c$  is the sample mean for treatment  $c$ .

The SSE is calculated:

$$SSE = \sum(x - \bar{x}_N)^2 + \sum(x - \bar{x}_W)^2 + \sum(x - \bar{x}_P)^2 + \sum(x - \bar{x}_B)^2$$

The detailed calculations of SS total and SSE for this example follow. To determine the values of SS total and SSE, we start by calculating the overall or grand mean. There are 22 observations and the total is 1,664, so the grand mean is 75.64.

$$\bar{x}_G = \frac{1,664}{22} = 75.64$$

|              | Northern | WTA   | Pocono | Branson | Total |
|--------------|----------|-------|--------|---------|-------|
|              | 94       | 75    | 70     | 68      |       |
|              | 90       | 68    | 73     | 70      |       |
|              | 85       | 77    | 76     | 72      |       |
|              | 80       | 83    | 78     | 65      |       |
|              |          | 88    | 80     | 74      |       |
|              |          |       | 68     | 65      |       |
|              |          |       | 65     |         |       |
| Column total | 349      | 391   | 510    | 414     | 1,664 |
| <i>n</i>     | 4        | 5     | 7      | 6       | 22    |
| Mean         | 87.25    | 78.20 | 72.86  | 69.00   | 75.64 |

Next we find the deviation of each observation from the grand mean, square those deviations, and sum this result for all 22 observations. For example, the first sampled passenger had a score of 94 and the overall or grand mean is 75.64. So  $(x - \bar{x}_G) = 94 - 75.64 = 18.36$ . For the last passenger,  $(x - \bar{x}_G) = 65 - 75.64 = -10.64$ . The calculations for all other passengers follow.

| Northern | WTA   | Pocono | Branson |
|----------|-------|--------|---------|
| 18.36    | -0.64 | -5.64  | -7.64   |
| 14.36    | -7.64 | -2.64  | -5.64   |
| 9.36     | 1.36  | 0.36   | -3.64   |
| 4.36     | 7.36  | 2.36   | -10.64  |
|          | 12.36 | 4.36   | -1.64   |
|          |       | -7.64  | -10.64  |
|          |       | -10.64 |         |

Then square each of these differences and sum all the values. Thus, for the first passenger:

$$(x - \bar{x}_G)^2 = (94 - 75.64)^2 = (18.36)^2 = 337.09$$

Finally, sum all the squared differences as formula (12-2) directs. Our SS total value is 1,485.10.

|       | Northern | WTA    | Pocono | Branson | Total    |
|-------|----------|--------|--------|---------|----------|
|       | 337.09   | 0.41   | 31.81  | 58.37   |          |
|       | 206.21   | 58.37  | 6.97   | 31.81   |          |
|       | 87.61    | 1.85   | 0.13   | 13.25   |          |
|       | 19.01    | 54.17  | 5.57   | 113.21  |          |
|       |          | 152.77 | 19.01  | 2.69    |          |
|       |          |        | 58.37  | 113.21  |          |
|       |          |        | 113.21 |         |          |
| Total | 649.92   | 267.57 | 235.07 | 332.54  | 1,485.10 |

To compute the term SSE, find the deviation between each observation and its treatment mean. In the example, the mean of the first treatment (that is, the passengers on Northern Airlines) is 87.25, found by  $\bar{x}_N = 349/4$ . The subscript *N* refers to Northern Airlines.

The first passenger rated Northern a 94, so  $(x - \bar{x}_N) = (94 - 87.25) = 6.75$ . The first passenger in the WTA group responded with a total score of 75, so  $(x - \bar{x}_W) = (75 - 78.20) = -3.2$ . The detail for all the passengers follows.

| Northern | WTA   | Pocono | Branson |
|----------|-------|--------|---------|
| 6.75     | -3.2  | -2.86  | -1      |
| 2.75     | -10.2 | 0.14   | 1       |
| -2.25    | -1.2  | 3.14   | 3       |
| -7.25    | 4.8   | 5.14   | -4      |
|          | 9.8   | 7.14   | 5       |
|          |       | -4.86  | -4      |
|          |       | -7.86  |         |

Each of these values is squared and then summed for all 22 observations. The four column totals can also be summed to find SSE. The values are shown in the following table.

|       | Northern | WTA    | Pocono | Branson | Total  |
|-------|----------|--------|--------|---------|--------|
|       | 45.5625  | 10.24  | 8.18   | 1       |        |
|       | 7.5625   | 104.04 | 0.02   | 1       |        |
|       | 5.0625   | 1.44   | 9.86   | 9       |        |
|       | 52.5625  | 23.04  | 26.42  | 16      |        |
|       |          | 96.04  | 50.98  | 25      |        |
|       |          |        | 23.62  | 16      |        |
|       |          |        | 61.78  |         |        |
| Total | 110.7500 | 234.80 | 180.86 | 68      | 594.41 |

So the SSE value is 594.41. That is,  $\Sigma(x - \bar{x}_c)^2 = 594.41$ .

Finally, we determine SST, the sum of the squares due to the treatments, by subtraction.

$$SST = SS \text{ total} - SSE \tag{12-4}$$

For this example:

$$SST = SS \text{ total} - SSE = 1,485.10 - 594.41 = 890.69.$$

To find the computed value of  $F$ , work your way across the ANOVA table. The degrees of freedom for the numerator and the denominator are the same as in step 4 on page 344 when we were finding the critical value of  $F$ . The term **mean square** is another expression for an estimate of the variance. The mean square for treatments is SST divided by its degrees of freedom. The result is the **mean square for treatments** and is written MST. Compute the **mean square error** in a similar fashion. To be precise, divide SSE by its degrees of freedom. To complete the process and find  $F$ , divide MST by MSE.

Insert the particular values of  $F$  into an ANOVA table and compute the value of  $F$  as follows.

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | $F$  |
|---------------------|----------------|--------------------|-------------|------|
| Treatments          | 890.69         | 3                  | 296.90      | 8.99 |
| Error               | 594.41         | 18                 | 33.02       |      |
| Total               | 1,485.10       | 21                 |             |      |

The computed value of  $F$  is 8.99, which is greater than the critical value of 5.09, so the null hypothesis is rejected.

**Step 6: Interpret the result.** We conclude the population means are not all equal. At this point, the results of the ANOVA only show that at least one pair of mean satisfaction scores are not the same among the four airlines. We cannot statistically show which airlines differ in satisfaction or which airlines have the highest or lowest satisfaction scores. The techniques for determining how the airlines differ are presented in the next section.

The calculations in the previous example/solution are tedious if the number of observations in each treatment is large. Many statistical software packages will perform the calculations and output the results. In the following illustration, Excel is used to calculate the descriptive statistics and ANOVA for the previous example/solution involving airlines and passenger ratings. There are some slight differences between the output and the previous calculations. These differences are due to rounding.

|    | A        | B   | C      | D       | E | F                    | G        | H   | I       | J        | K       | L      | M |
|----|----------|-----|--------|---------|---|----------------------|----------|-----|---------|----------|---------|--------|---|
| 1  | Northern | WTA | Pocono | Branson |   | Anova: Single Factor |          |     |         |          |         |        |   |
| 2  | 94       | 75  | 70     | 68      |   |                      |          |     |         |          |         |        |   |
| 3  | 90       | 68  | 73     | 70      |   | SUMMARY              |          |     |         |          |         |        |   |
| 4  | 85       | 77  | 76     | 72      |   | Groups               | Count    | Sum | Average | Variance |         |        |   |
| 5  | 80       | 83  | 78     | 65      |   | Northern             | 4        | 349 | 87.250  | 36.917   |         |        |   |
| 6  |          | 88  | 80     | 74      |   | WTA                  | 5        | 391 | 78.200  | 58.700   |         |        |   |
| 7  |          |     | 68     | 65      |   | Pocono               | 7        | 510 | 72.857  | 30.143   |         |        |   |
| 8  |          |     | 65     |         |   | Branson              | 6        | 414 | 69.000  | 13.600   |         |        |   |
| 9  |          |     |        |         |   |                      |          |     |         |          |         |        |   |
| 10 |          |     |        |         |   | ANOVA                |          |     |         |          |         |        |   |
| 11 |          |     |        |         |   | Source of Variation  | SS       | df  | MS      | F        | P-value | F crit |   |
| 12 |          |     |        |         |   | Between Groups       | 890.684  | 3   | 296.895 | 8.99     | 0.0007  | 3.160  |   |
| 13 |          |     |        |         |   | Within Groups        | 594.407  | 18  | 33.023  |          |         |        |   |
| 14 |          |     |        |         |   | Total                | 1485.091 | 21  |         |          |         |        |   |
| 15 |          |     |        |         |   |                      |          |     |         |          |         |        |   |
| 16 |          |     |        |         |   |                      |          |     |         |          |         |        |   |

Source: Microsoft Excel

Notice Excel uses the term “Between Groups” for treatments and “Within Groups” for error. However, they have the same meanings. The  $p$ -value is .0007. This is the probability of finding a value of the test statistic this large or larger when the null hypothesis is true. To put it another way, it is the likelihood of calculating an  $F$ -value larger than 8.99 with 3 degrees of freedom in the numerator and 18 degrees of freedom in the denominator. So when we reject the null hypothesis in this instance, there is a very small likelihood of committing a Type I error!

## SELF-REVIEW 12-2



Citrus Clean is a new all-purpose cleaner being test-marketed by placing displays in three different locations within various supermarkets. The number of 12-ounce bottles sold from each location within the supermarket is reported below.

| Near Bread | Near Beer | With Cleaners |
|------------|-----------|---------------|
| 18         | 12        | 26            |
| 14         | 18        | 28            |
| 19         | 10        | 30            |
| 17         | 16        | 32            |

At the .05 significance level, is there a difference in the mean number of bottles sold at the three locations?

- State the null hypothesis and the alternate hypothesis.
- What is the decision rule?
- Compute the values of SS total, SST, and SSE.
- Develop an ANOVA table.
- What is your decision regarding the null hypothesis?

## EXERCISES

7. The following are four observations collected from each of three treatments. Test the hypothesis that the treatment means are equal. Use the .05 significance level.

| Treatment 1 | Treatment 2 | Treatment 3 |
|-------------|-------------|-------------|
| 8           | 3           | 3           |
| 6           | 2           | 4           |
| 10          | 4           | 5           |
| 9           | 3           | 4           |

- State the null and the alternate hypotheses.
  - What is the decision rule?
  - Compute SST, SSE, and SS total.
  - Complete an ANOVA table.
  - State your decision regarding the null hypothesis.
8. The following are six observations collected from treatment 1, four observations collected from treatment 2, and five observations collected from treatment 3. Test the hypothesis at the .05 significance level that the treatment means are equal.

| Treatment 1 | Treatment 2 | Treatment 3 |
|-------------|-------------|-------------|
| 9           | 13          | 10          |
| 7           | 20          | 9           |
| 11          | 14          | 15          |
| 9           | 13          | 14          |
| 12          |             | 15          |
| 10          |             |             |

- State the null and the alternate hypotheses.
  - What is the decision rule?
  - Compute SST, SSE, and SS total.
  - Complete an ANOVA table.
  - State your decision regarding the null hypothesis.
9. **FILE** A real estate developer is considering investing in a shopping mall on the outskirts of Atlanta, Georgia. Three parcels of land are being evaluated. Of particular importance is the income in the area surrounding the proposed mall. A random sample of four families is selected near each proposed mall. Following are the sample results. At the .05 significance level, can the developer conclude there is a difference in the mean income? Use the usual six-step hypothesis testing procedure.

| Southwyck Area (\$000) | Franklin Park (\$000) | Old Orchard (\$000) |
|------------------------|-----------------------|---------------------|
| 64                     | 74                    | 75                  |
| 68                     | 71                    | 80                  |
| 70                     | 69                    | 76                  |
| 60                     | 70                    | 78                  |

10. **FILE** The manager of a computer software company wishes to study the number of hours per week senior executives by type of industry spend at their desktop computers. The manager selected a sample of five executives from each of three industries. At the .05 significance level, can she conclude there is a difference in the mean number of hours spent per week by industry?

| Banking | Retail | Insurance |
|---------|--------|-----------|
| 32      | 28     | 30        |
| 30      | 28     | 28        |
| 30      | 26     | 26        |
| 32      | 28     | 28        |
| 30      | 30     | 30        |

**LO12-3**

Use confidence intervals to test and interpret differences between pairs of population means.

## INFERENCES ABOUT PAIRS OF TREATMENT MEANS

Suppose we carry out the ANOVA procedure, make the decision to reject the null hypothesis, and conclude that all the treatment means are not the same. Sometimes we may be satisfied with this conclusion, but in other instances we may want to know which treatment means differ. This section provides the details for this analysis.

Recall in the previous example/solution regarding airline passenger ratings, we concluded that there was a difference in the treatment means. That is, the null hypothesis was rejected and the alternate hypothesis accepted. The conclusion is that at least one of the airlines' mean level of satisfaction is different from the others. Now, the question is which of the four airlines differ?

Several procedures are available to answer this question. The simplest is through the use of confidence intervals, that is, formula (9–2). From the computer output of the example on page 348, the sample mean score for those passengers rating Northern's service is 87.25, and for those rating Branson's service, the sample mean score is 69.00. Is there enough disparity to justify the conclusion that there is a significant difference in the mean satisfaction scores of the two airlines?

The  $t$  distribution, described in Chapters 10 and 11, is used as the basis for this test. Recall that one of the assumptions of ANOVA is that the population variances are the same for all treatments. This common population value is the **mean square error**, or MSE, and is determined by  $SSE/(n - k)$ . A confidence interval for the difference between two populations is found by:

$$\text{CONFIDENCE INTERVAL FOR THE DIFFERENCE IN TREATMENT MEANS} \quad (\bar{x}_1 - \bar{x}_2) \pm t \sqrt{\text{MSE} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (12-5)$$

where:

$\bar{x}_1$  is the mean of the first sample.

$\bar{x}_2$  is the mean of the second sample.

$t$  is obtained from Appendix B.5. The degrees of freedom are equal to  $n - k$ .

MSE is the mean square error term obtained from the ANOVA table [ $SSE/(n - k)$ ].

$n_1$  is the number of observations in the first sample.

$n_2$  is the number of observations in the second sample.

How do we decide whether there is a difference in the treatment means? If the confidence interval includes zero, there is *not* a difference between the treatment means. For example, if the left endpoint of the confidence interval has a negative sign and the right endpoint has a positive sign, the interval includes zero and the two means do not differ. So if we develop a confidence interval from formula (12–5) and find the difference in the sample means was 5.00—that is, if  $\bar{x}_1 - \bar{x}_2 = 5$  and

$t \sqrt{\text{MSE} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} = 12$ —the confidence interval would range from  $-7.00$  up to  $17.00$ .

To put it in symbols:

$$(\bar{x}_1 - \bar{x}_2) \pm t \sqrt{\text{MSE} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} = 5.00 \pm 12.00 = -7.00 \text{ up to } 17.00$$

Note that zero is in this interval. Therefore, we conclude that there is no significant difference in the selected treatment means.

On the other hand, if the endpoints of the confidence interval have the same sign, this indicates that the treatment means differ. For example, if  $\bar{x}_1 - \bar{x}_2 = -0.35$  and

$t\sqrt{\text{MSE}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 0.25$ , the confidence interval would range from  $-0.60$  up to  $-0.10$ . Because  $-0.60$  and  $-0.10$  have the same sign, both negative, zero is not in the interval and we conclude that these treatment means differ.

Using the previous airline example, let us compute the confidence interval for the difference between the mean scores of passengers on Northern and Branson. With a 95% level of confidence, the endpoints of the confidence interval are 10.457 and 26.043.

$$(\bar{x}_N - \bar{x}_B) \pm t\sqrt{\text{MSE}\left(\frac{1}{n_N} + \frac{1}{n_B}\right)} = (87.25 - 69.00) \pm 2.101\sqrt{33.023\left(\frac{1}{4} + \frac{1}{6}\right)} = 18.25 \pm 7.793$$

where:

$\bar{x}_N$  is 87.25.

$\bar{x}_B$  is 69.00.

$t$  is 2.101: from Appendix B.5 with  $(n - k) = 22 - 4 = 18$  degrees of freedom.

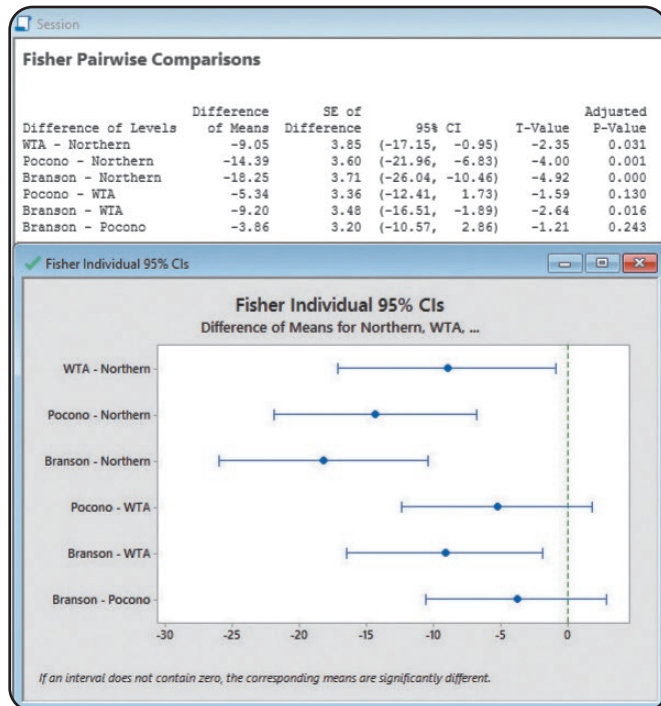
MSE is 33.023: from the ANOVA table with  $\text{SSE}/(n - k) = 594.4/18$ .

$n_N$  is 4.

$n_B$  is 6.

The 95% confidence interval ranges from 10.457 up to 26.043. Both endpoints are positive; hence, we can conclude these treatment means differ significantly. That is, passengers on Northern Airlines rated service significantly different from those on Branson Airlines.

The confidence intervals for the differences between each pair of means can be obtained directly using statistical software. The following confidence intervals were computed using the one-way ANOVA in Minitab. Statistical software, such as Minitab, offers a variety of methods to control Type I error when making multiple comparisons. The following analysis used Fisher's method to compare means.



Source: Minitab



The output shows the confidence intervals for the difference between each pair of treatment means. The first row shows the confidence interval that compares WTA and Northern. It shows a confidence interval that does not include zero. It also shows the  $p$ -value for a hypothesis test that the means of WTA and Northern are equal. The hypothesis is rejected because a  $p$ -value of 0.031 is less than an assumed  $\alpha$  of 0.05. Both results indicate that the WTA and Northern means are significantly different. Reviewing the entire table, only two pairs of means are not significantly different: Pocono and WTA, and Branson and Pocono. All other confidence intervals do not include zero and have  $p$ -values less than 0.05. Therefore, all other pairs of means are significantly different.

The graphic illustrates the results of the confidence interval analysis. Each confidence interval is represented by its endpoints and treatment mean. Note that a difference of zero is illustrated with the vertical dotted line. Two of the intervals include zero, Pocono and WTA, and Branson and Pocono. The others do not include zero so the means are significantly different. The following pairs of means are different: WTA and Northern, Pocono and Northern, Branson and Northern, and Branson and WTA.

We should emphasize that this investigation is a step-by-step process. The initial step is to conduct the ANOVA test. Only if the null hypothesis that the treatment means are equal is rejected, should any analysis of the individual treatment means be attempted.

## SELF-REVIEW 12-3



The following data are the semester tuition charges (\$000) for a sample of five private colleges in the Northeast region of the United States, four in the Southeast region, and five in the West region. At the .05 significance level, can we conclude there is a difference in the mean tuition rates for the various regions?

| Northeast (\$000) | Southeast (\$000) | West (\$000) |
|-------------------|-------------------|--------------|
| 40                | 38                | 37           |
| 41                | 39                | 38           |
| 42                | 40                | 36           |
| 40                | 38                | 37           |
| 42                |                   | 36           |

- State the null and the alternate hypotheses.
- What is the decision rule?
- Develop an ANOVA table. What is the value of the test statistic?
- What is your decision regarding the null hypothesis?
- Could there be a significant difference between the mean tuition in the Northeast and that of the West? If so, develop a 95% confidence interval for that difference.

## EXERCISES

11. **FILE** The following are three observations collected from treatment 1, five observations collected from treatment 2, and four observations collected from treatment 3. Test the hypothesis that the treatment means are equal at the .05 significance level.

| Treatment 1 | Treatment 2 | Treatment 3 |
|-------------|-------------|-------------|
| 8           | 3           | 3           |
| 11          | 2           | 4           |
| 10          | 1           | 5           |
|             | 3           | 4           |
|             | 2           |             |

- a. State the null hypothesis and the alternate hypothesis.
  - b. What is the decision rule?
  - c. Compute SST, SSE, and SS total.
  - d. Complete an ANOVA table.
  - e. State your decision regarding the null hypothesis.
  - f. If  $H_0$  is rejected, can we conclude that treatment 1 and treatment 2 differ? Use the 95% level of confidence.
12. **FILE** The following are six observations collected from treatment 1, ten observations collected from treatment 2, and eight observations collected from treatment 3. Test the hypothesis that the treatment means are equal at the .05 significance level.

| Treatment 1 | Treatment 2 | Treatment 3 |
|-------------|-------------|-------------|
| 3           | 9           | 6           |
| 2           | 6           | 3           |
| 5           | 5           | 5           |
| 1           | 6           | 5           |
| 3           | 8           | 5           |
| 1           | 5           | 4           |
|             | 4           | 1           |
|             | 7           | 5           |
|             | 6           |             |
|             | 4           |             |

- a. State the null hypothesis and the alternate hypothesis.
  - b. What is the decision rule?
  - c. Compute SST, SSE, and SS total.
  - d. Complete an ANOVA table.
  - e. State your decision regarding the null hypothesis.
  - f. If  $H_0$  is rejected, can we conclude that treatment 2 and treatment 3 differ? Use the 95% level of confidence.
13. A senior accounting major at Midsouth State University has job offers from four CPA firms. To explore the offers further, she asked a sample of recent trainees how many months each worked for the firm before receiving a raise in salary. The sample information is submitted to Minitab with the following results:

| Analysis of Variance |    |       |       |      |       |
|----------------------|----|-------|-------|------|-------|
| Source               | DF | SS    | MS    | F    | P     |
| Factor               | 3  | 32.33 | 10.78 | 2.36 | 0.133 |
| Error                | 10 | 45.67 | 4.57  |      |       |
| Total                | 13 | 78.00 |       |      |       |

At the .05 level of significance, is there a difference in the mean number of months before a raise was granted among the four CPA firms?

14. **FILE** A stock analyst wants to determine whether there is a difference in the mean return on equity for three types of stock: utility, retail, and banking stocks. The following output is obtained:

## Analysis of Variance

| Source | DF | Adj SS  | Adj MS  | F-Value | P-Value |
|--------|----|---------|---------|---------|---------|
| Factor | 2  | 303.697 | 151.848 | 5.50    | 0.0202  |
| Error  | 12 | 331.381 | 27.615  |         |         |
| Total  | 14 | 635.077 |         |         |         |

## Means

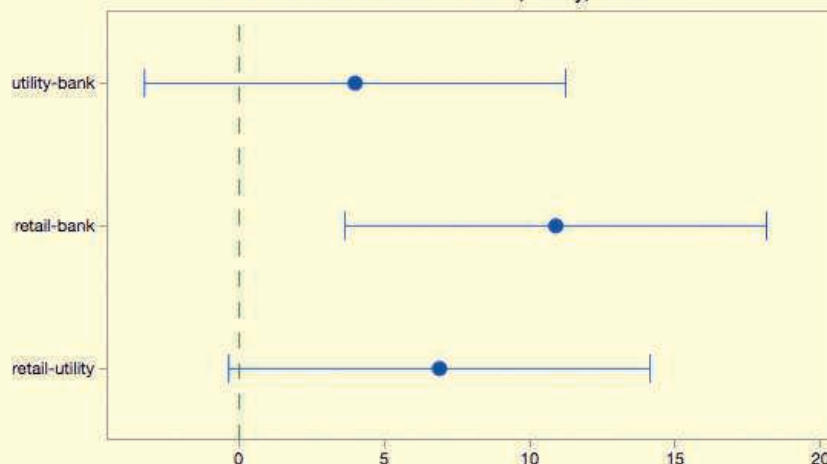
| Factor  | N | Mean   | StDev  | 95% CI            |
|---------|---|--------|--------|-------------------|
| bank    | 5 | 5.318  | 3.904  | (0.198, 10.438)   |
| utility | 5 | 9.3160 | 1.5376 | (4.1956, 14.4364) |
| retail  | 5 | 16.212 | 8.077  | (11.092, 21.332)  |

Pooled StDev = 5.25500

## Fisher Individual Tests for Differences of Means

| Difference of Levels | Difference of Means | SE of Difference | 95% CI           | T-Value | Adjusted P-Value |
|----------------------|---------------------|------------------|------------------|---------|------------------|
| utility-bank         | 3.998               | 3.324            | (-3.243, 11.239) | 1.20    | 0.2522           |
| retail-bank          | 10.894              | 3.324            | (3.653, 18.135)  | 3.28    | 0.0066           |
| retail-utility       | 6.896               | 3.324            | (-0.345, 14.137) | 2.07    | 0.0602           |

Fisher Individual 95% CIs  
Differences of Means for bank, utility, retail



If an interval does not contain 0, the corresponding means are significantly different.

- Using the .05 level of significance, is there a difference in the mean return on equity among the three types of stock?
- Can the analyst conclude there is a difference between the mean return on equity for utility and retail stocks? For utility and banking stocks? For banking and retail stocks? Explain.

## CHAPTER SUMMARY

- The characteristics of the  $F$  distribution are:
  - It is continuous.
  - Its values cannot be negative.
  - It is positively skewed.
  - There is a family of  $F$  distributions. Each time the degrees of freedom in either the numerator or the denominator change, a new distribution is created.

- II. The  $F$  distribution is used to test whether two population variances are the same.
  - A. The sampled populations must follow the normal distribution.
  - B. The larger of the two sample variances is placed in the numerator, forcing the ratio to be at least 1.00.
  - C. The value of  $F$  is computed using the following equation:

$$F = \frac{s_1^2}{s_2^2} \tag{12-1}$$

- III. A one-way ANOVA is used to compare several treatment means.
  - A. A treatment is a source of variation.
  - B. The assumptions underlying ANOVA are:
    - 1. The samples are from populations that follow the normal distribution.
    - 2. The populations have equal standard deviations.
    - 3. The populations are independent.
  - C. The information for finding the value of  $F$  is summarized in an ANOVA table.
    - 1. The formula for SS total, the sum of squares total, is:

$$SS \text{ total} = \Sigma(x - \bar{x}_G)^2 \tag{12-2}$$

- 2. The formula for SSE, the sum of squares error, is:

$$SSE = \Sigma(x - \bar{x}_c)^2 \tag{12-3}$$

- 3. The formula for the SST, the sum of squares treatment, is found by subtraction.

$$SST = SS \text{ total} - SSE \tag{12-4}$$

- 4. This information is summarized in the following ANOVA table and the value of  $F$  is determined.

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square         | $F$       |
|---------------------|----------------|--------------------|---------------------|-----------|
| Treatments          | SST            | $k - 1$            | $SST/(k - 1) = MST$ | $MST/MSE$ |
| Error               | SSE            | $n - k$            | $SSE/(n - k) = MSE$ |           |
| Total               | SS total       | $n - 1$            |                     |           |

- IV. If a null hypothesis of equal treatment means is rejected, we can identify the pairs of means that differ from the following confidence interval.

$$(\bar{x}_1 - \bar{x}_2) \pm t \sqrt{MSE \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \tag{12-5}$$

**PRONUNCIATION KEY**

| SYMBOL   | MEANING                  | PRONUNCIATION |
|----------|--------------------------|---------------|
| SS total | Sum of squares total     | S S total     |
| SST      | Sum of squares treatment | S S T         |
| SSE      | Sum of squares error     | S S E         |
| MSE      | Mean square error        | M S E         |

**CHAPTER EXERCISES**

- 15. A real estate agent in the coastal area of Georgia wants to compare the variation in the selling price of homes on the oceanfront with those one to three blocks from the ocean. A sample of 21 oceanfront homes sold within the last year revealed the standard deviation of the selling prices was \$45,600. A sample of 18 homes, also sold within the last

year, that were one to three blocks from the ocean revealed that the standard deviation was \$21,330. At the .01 significance level, can we conclude that there is more variation in the selling prices of the oceanfront homes?

16. One variable that Google uses to rank pages on the Internet is page speed, the time it takes for a web page to load into your browser. A source for women’s clothing is redesigning their page to improve the images that show its products and to reduce its load time. The new page is clearly faster, but initial tests indicate there is more variation in the time to load. A sample of 16 different load times showed that the standard deviation of the load time was 22 hundredths of a second for the new page and 12 hundredths of a second for the current page. At the .05 significance level, can we conclude that there is more variation in the load time of the new page?
17. **FILE** There are two Chevrolet dealers in Jamestown, New York. The mean monthly sales at Sharkey Chevy and Dave White Chevrolet are about the same. However, Tom Sharkey, the owner of Sharkey Chevy, believes his sales are more consistent. Below are the numbers of new cars sold at Sharkey in the last 7 months and for the last 8 months at Dave White. Do you agree with Mr. Sharkey? Use the .01 significance level.

|            |    |    |    |    |    |    |    |     |
|------------|----|----|----|----|----|----|----|-----|
| Sharkey    | 98 | 78 | 54 | 57 | 68 | 64 | 70 |     |
| Dave White | 75 | 81 | 81 | 30 | 82 | 46 | 58 | 101 |

18. Random samples of five were selected from each of three populations. The sum of squares total was 100. The sum of squares due to the treatments was 40.
  - a. Set up the null hypothesis and the alternate hypothesis.
  - b. What is the decision rule? Use the .05 significance level.
  - c. Create the ANOVA table. What is the value of  $F$ ?
  - d. What is your decision regarding the null hypothesis?
19. In an ANOVA table, the MSE is equal to 10. Random samples of six were selected from each of four populations, where the sum of squares total was 250.
  - a. Set up the null hypothesis and the alternate hypothesis.
  - b. What is the decision rule? Use the .05 significance level.
  - c. Create the ANOVA table. What is the value of  $F$ ?
  - d. What is your decision regarding the null hypothesis?
20. The following is a partial ANOVA table.

| Source    | Sum of Squares | $df$ | Mean Square | $F$ |
|-----------|----------------|------|-------------|-----|
| Treatment |                | 2    |             |     |
| Error     |                |      | 20          |     |
| Total     | 500            | 11   |             |     |

Complete the table and answer the following questions. Use the .05 significance level.

- a. How many treatments are there?
- b. What is the total sample size?
- c. What is the critical value of  $F$ ?
- d. Write out the null and alternate hypotheses.
- e. What is your conclusion regarding the null hypothesis?
21. **FILE** A consumer organization wants to know whether there is a difference in the price of a particular toy at three different types of stores. The price of the toy was checked in a sample of five discount stores, five variety stores, and five department stores. The results are shown below. Use the .05 significance level.

| Discount | Variety | Department |
|----------|---------|------------|
| \$12     | \$15    | \$19       |
| 13       | 17      | 17         |
| 14       | 14      | 16         |
| 12       | 18      | 20         |
| 15       | 17      | 19         |

22. **FILE** Jacob Lee is a frequent traveler between Los Angeles and San Diego. For the past month, he wrote down the flight times in minutes on three different airlines. The results are:

| Goust | Jet Red | Cloudtran |
|-------|---------|-----------|
| 51    | 50      | 52        |
| 51    | 53      | 55        |
| 52    | 52      | 60        |
| 42    | 62      | 64        |
| 51    | 53      | 61        |
| 57    | 49      | 49        |
| 47    | 50      | 49        |
| 47    | 49      |           |
| 50    | 58      |           |
| 60    | 54      |           |
| 54    | 51      |           |
| 49    | 49      |           |
| 48    | 49      |           |
| 48    | 50      |           |

- a. Use the .05 significance level and the six-step hypothesis-testing process to check if there is a difference in the mean flight times among the three airlines.
- b. Develop a 95% confidence interval for the difference in the means between Goust and Cloudtran.
23. **FILE** The City of Maumee comprises four districts. Chief of Police Andy North wants to determine whether there is a difference in the mean number of crimes committed among the four districts. He examined the records from six randomly selected days and recorded the number of crimes. At the .05 significance level, can Chief North conclude that there is a difference in the mean number of crimes among the four districts?

| Number of Crimes |            |          |            |
|------------------|------------|----------|------------|
| Rec Center       | Key Street | Monclova | Whitehouse |
| 13               | 21         | 12       | 16         |
| 15               | 13         | 14       | 17         |
| 14               | 18         | 15       | 18         |
| 15               | 19         | 13       | 15         |
| 14               | 18         | 12       | 20         |
| 15               | 19         | 15       | 18         |

24. **FILE** A study of the effect of television commercials on 12-year-old children measured their attention span, in seconds. The commercials were for clothes, food, and toys. At the .05 significance level, is there a difference in the mean attention span of the children for the various commercials? Are there significant differences between pairs of means? Would you recommend dropping one of the three commercial types?

| Clothes | Food | Toys |
|---------|------|------|
| 26      | 45   | 60   |
| 21      | 48   | 51   |
| 43      | 43   | 43   |
| 35      | 53   | 54   |
| 28      | 47   | 63   |
| 31      | 42   | 53   |
| 17      | 34   | 48   |
| 31      | 43   | 58   |
| 20      | 57   | 47   |
|         | 47   | 51   |
|         | 44   | 51   |
|         | 54   |      |

25. **FILE** When only two treatments are involved, ANOVA and the Student's  $t$  test (Chapter 11) result in the same conclusions. Also, for computed test statistics,  $t^2 = F$ . To demonstrate this relationship, use the following example. Fourteen randomly selected students enrolled in a history course were divided into two groups, one consisting of 6 students who took the course in the normal lecture format. The other group of 8 students took the course in a distance format. At the end of the course, each group was examined with a 50-item test. The following is a list of the number correct for each of the two groups.

| Traditional Lecture | Distance |
|---------------------|----------|
| 37                  | 50       |
| 35                  | 46       |
| 41                  | 49       |
| 40                  | 44       |
| 35                  | 41       |
| 34                  | 42       |
|                     | 45       |
|                     | 43       |

- a. Using analysis of variance techniques, test  $H_0$  that the two mean test scores are equal;  $\alpha = .05$ .
- b. Using the  $t$  test from Chapter 11, compute  $t$ .
- c. Interpret the results.
26. There are four auto body shops in Bangor, Maine, and all claim to promptly repair cars. To check if there is any difference in repair times, customers are randomly selected from each repair shop and their repair times in days are recorded. The output from a statistical software package is:

| Summary     |             |      |          |          |
|-------------|-------------|------|----------|----------|
| Groups      | Sample Size | Sum  | Average  | Variance |
| Body Shop A | 3           | 15.4 | 5.133333 | 0.323333 |
| Body Shop B | 4           | 32   | 8        | 1.433333 |
| Body Shop C | 5           | 25.2 | 5.04     | 0.748    |
| Body Shop D | 4           | 25.9 | 6.475    | 0.595833 |

| ANOVA               |                 |           |          |          |          |
|---------------------|-----------------|-----------|----------|----------|----------|
| Source of Variation | SS              | df        | MS       | F        | p-value  |
| Between Groups      | 23.37321        | 3         | 7.791069 | 9.612506 | 0.001632 |
| Within Groups       | <u>9.726167</u> | <u>12</u> | 0.810514 |          |          |
| Total               | 33.09938        | 15        |          |          |          |

Is there evidence to suggest a difference in the mean repair times at the four body shops? Use the .05 significance level.

27. The fuel efficiencies for a sample of 27 compact, midsize, and large cars are entered into a statistical software package. Analysis of variance is used to investigate if there is a difference in the mean miles per gallon of the three car sizes. What do you conclude? Use the .01 significance level.

| Summary |             |       |          |          |
|---------|-------------|-------|----------|----------|
| Groups  | Sample Size | Sum   | Average  | Variance |
| Compact | 12          | 268.3 | 22.35833 | 9.388106 |
| Midsize | 9           | 172.4 | 19.15556 | 7.315278 |
| Large   | 6           | 100.5 | 16.75    | 7.303    |

Additional results are shown below.

| ANOVA               |                 |           |          |          |          |
|---------------------|-----------------|-----------|----------|----------|----------|
| Source of Variation | SS              | df        | MS       | F        | p-value  |
| Between Groups      | 136.4803        | 2         | 68.24014 | 8.258752 | 0.001866 |
| Within Groups       | <u>198.3064</u> | <u>24</u> | 8.262766 |          |          |
| Total               | 334.7867        | 26        |          |          |          |

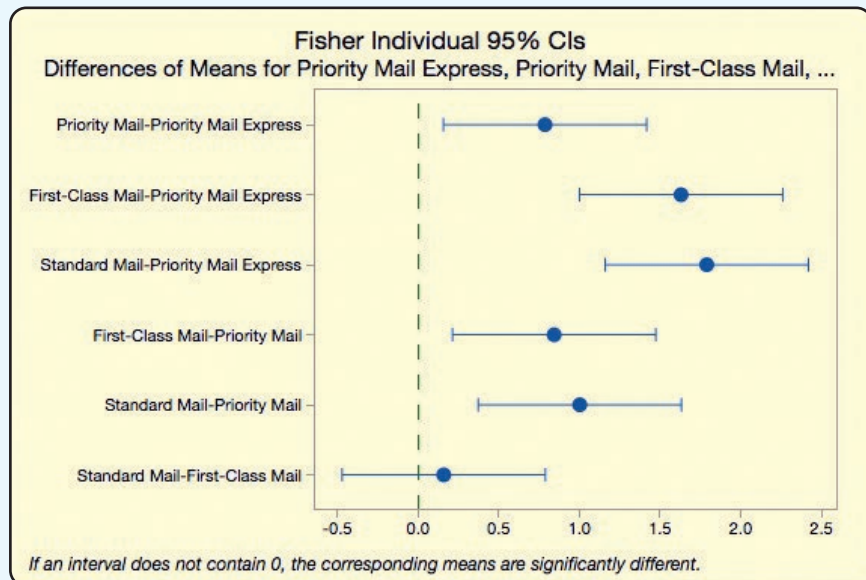
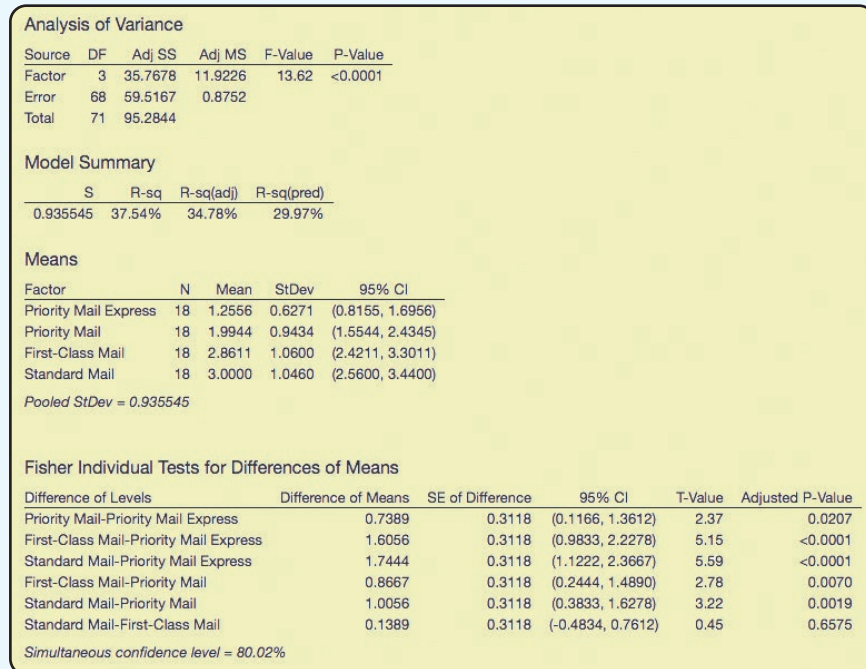
28. Three assembly lines are used to produce a certain component for an airliner. To examine the production rate, a random sample of six hourly periods is chosen for each assembly line and the number of components produced during these periods for each line is recorded. The output from a statistical software package is:

| Summary |             |     |          |          |
|---------|-------------|-----|----------|----------|
| Groups  | Sample Size | Sum | Average  | Variance |
| Line A  | 6           | 250 | 41.66667 | 0.266667 |
| Line B  | 6           | 260 | 43.33333 | 0.666667 |
| Line C  | 6           | 249 | 41.5     | 0.7      |

| ANOVA               |                 |           |          |          |          |
|---------------------|-----------------|-----------|----------|----------|----------|
| Source of Variation | SS              | df        | MS       | F        | p-value  |
| Between Groups      | 12.33333        | 2         | 6.166667 | 11.32653 | 0.001005 |
| Within Groups       | <u>8.166667</u> | <u>15</u> | 0.544444 |          |          |
| Total               | 20.5            | 17        |          |          |          |



- a. Use a .01 level of significance to test if there is a difference in the mean production of the three assembly lines.
  - b. Develop a 99% confidence interval for the difference in the means between Line B and Line C.
29. **FILE** The postal service sorts mail as Priority Mail Express, Priority Mail, First-Class Mail, or Standard Mail. Over a period of 3 weeks, 18 of each type were mailed from the Network Distribution Center in Atlanta, Georgia, to Des Moines, Iowa. The total delivery time in days was recorded. Minitab was used to perform the ANOVA. The results follow:



Using the ANOVA results, compare the average delivery times of the four different types of mail.

30. **FILE** To prevent spam from entering your e-mail inbox, you use a filter. You are interested in knowing if the number of spam e-mails differs by day of the week. The number of spam e-mails by day of week is counted and recorded. Minitab is used to perform the data analysis. Here are the results:

| Analysis of Variance |    |         |         |         |         |
|----------------------|----|---------|---------|---------|---------|
| Source               | DF | Adj SS  | Adj MS  | F-Value | P-Value |
| Factor               | 6  | 2266.92 | 377.820 | 9.29    | <0.0001 |
| Error                | 48 | 1952.43 | 40.676  |         |         |
| Total                | 54 | 4219.35 |         |         |         |

| Means     |    |        |       |                  |
|-----------|----|--------|-------|------------------|
| Factor    | N  | Mean   | StDev | 95% CI           |
| Monday    | 10 | 76.400 | 4.452 | (72.345, 80.455) |
| Tuesday   | 9  | 60.444 | 4.333 | (56.170, 64.719) |
| Wednesday | 7  | 74.286 | 4.231 | (69.439, 79.132) |
| Thursday  | 8  | 60.750 | 6.251 | (56.216, 65.284) |
| Friday    | 8  | 74.000 | 7.838 | (69.466, 78.534) |
| Saturday  | 5  | 64.000 | 7.036 | (58.265, 69.735) |
| Sunday    | 8  | 69.125 | 9.372 | (64.591, 73.659) |

Pooled StDev = 6.37774

| Grouping Information Using the Fisher LSD Method and 95% Confidence |    |        |          |  |
|---|----|--------|----------|--|
| Factor  | N  | Mean   | Grouping |  |
| Monday  | 10 | 76.400 | A        |  |
| Wednesday   | 7  | 74.286 | A B      |  |
| Friday  | 8  | 74.000 | A B      |  |
| Sunday  | 8  | 69.125 | B C      |  |
| Saturday  | 5  | 64.000 | C D      |  |
| Thursday  | 8  | 60.750 | D        |  |
| Tuesday   | 9  | 60.444 | D        |  |

Means that do not share a letter are significantly different.

| Fisher Individual Tests for Differences of Means |                     |                  |                    |         |                  |  |
|--|---------------------|------------------|--------------------|---------|------------------|--|
| Difference of Levels                             | Difference of Means | SE of Difference | 95% CI             | T-Value | Adjusted P-Value |  |
| Tuesday-Monday                                   | -15.956             | 2.930            | (-21.847, -10.064) | -5.44   | <0.0001          |  |
| Wednesday-Monday                                 | -2.114              | 3.143            | (-8.434, 4.205)    | -0.67   | 0.5044           |  |
| Thursday-Monday                                  | -15.650             | 3.025            | (-21.733, -9.567)  | -5.17   | <0.0001          |  |
| Friday-Monday                                    | -2.400              | 3.025            | (-8.483, 3.683)    | -0.79   | 0.4315           |  |
| Saturday-Monday                                  | -12.400             | 3.493            | (-19.424, -5.376)  | -3.55   | 0.0009           |  |
| Sunday-Monday                                    | -7.275              | 3.025            | (-13.358, -1.192)  | -2.40   | 0.0201           |  |
| Wednesday-Tuesday                                | 13.841              | 3.214            | (7.379, 20.304)    | 4.31    | <0.0001          |  |
| Thursday-Tuesday                                 | 0.306               | 3.099            | (-5.925, 6.537)    | 0.10    | 0.9219           |  |
| Friday-Tuesday                                   | 13.556              | 3.099            | (7.325, 19.787)    | 4.37    | <0.0001          |  |
| Saturday-Tuesday                                 | 3.556               | 3.557            | (-3.597, 10.708)   | 1.00    | 0.3226           |  |
| Sunday-Tuesday                                   | 8.681               | 3.099            | (2.450, 14.912)    | 2.80    | 0.0073           |  |
| Thursday-Wednesday                               | -13.536             | 3.301            | (-20.172, -6.899)  | -4.10   | 0.0002           |  |
| Friday-Wednesday                                 | -0.286              | 3.301            | (-6.922, 6.351)    | -0.09   | 0.9314           |  |
| Saturday-Wednesday                               | -10.286             | 3.734            | (-17.794, -2.777)  | -2.75   | 0.0083           |  |
| Sunday-Wednesday                                 | -5.161              | 3.301            | (-11.797, 1.476)   | -1.56   | 0.1245           |  |
| Friday-Thursday                                  | 13.250              | 3.189            | (6.838, 19.662)    | 4.16    | 0.0001           |  |
| Saturday-Thursday                                | 3.250               | 3.636            | (-4.060, 10.560)   | 0.89    | 0.3759           |  |
| Sunday-Thursday                                  | 8.375               | 3.189            | (1.963, 14.787)    | 2.63    | 0.0115           |  |
| Saturday-Friday                                  | -10.000             | 3.636            | (-17.310, -2.690)  | -2.75   | 0.0084           |  |
| Sunday-Friday                                    | -4.875              | 3.189            | (-11.287, 1.537)   | -1.53   | 0.1329           |  |
| Sunday-Saturday                                  | 5.125               | 3.636            | (-2.185, 12.435)   | 1.41    | 0.1651           |  |

Simultaneous confidence level = 57.84%

Using the ANOVA results, compare the average number of spam e-mails for each day of the week.

31. **FILE** Listed below are the weights (in grams) of a sample of M&M's Plain candies, classified according to color. Use a statistical software system to determine whether there is a difference in the mean weights of candies of different colors. Use the .05 significance level.

| Red   | Orange | Yellow | Brown | Tan   | Green |
|-------|--------|--------|-------|-------|-------|
| 0.946 | 0.902  | 0.929  | 0.896 | 0.845 | 0.935 |
| 1.107 | 0.943  | 0.960  | 0.888 | 0.909 | 0.903 |
| 0.913 | 0.916  | 0.938  | 0.906 | 0.873 | 0.865 |
| 0.904 | 0.910  | 0.933  | 0.941 | 0.902 | 0.822 |
| 0.926 | 0.903  | 0.932  | 0.838 | 0.956 | 0.871 |
| 0.926 | 0.901  | 0.899  | 0.892 | 0.959 | 0.905 |
| 1.006 | 0.919  | 0.907  | 0.905 | 0.916 | 0.905 |
| 0.914 | 0.901  | 0.906  | 0.824 | 0.822 | 0.852 |
| 0.922 | 0.930  | 0.930  | 0.908 |       | 0.965 |
| 1.052 | 0.883  | 0.952  | 0.833 |       | 0.898 |
| 0.903 |        | 0.939  |       |       |       |
| 0.895 |        | 0.940  |       |       |       |
|       |        | 0.882  |       |       |       |
|       |        | 0.906  |       |       |       |

32. There are four radio stations in Midland. The stations have different formats (hard rock, classical, country/western, and easy listening), but each is concerned with the number of minutes of music played per hour. From a sample of 10 randomly selected hours from each station, the sum of squared differences between each observation and the mean for its respective radio station,  $\sum(x - \bar{x}_c)^2$ , are:

|                    |        |                          |        |
|--------------------|--------|--------------------------|--------|
| Hard rock station: | 126.29 | Country/western station: | 166.79 |
| Classical station: | 233.34 | Easy listening station:  | 77.57  |

The total sum of squares for the data is:  $SS_{\text{total}} = 1,099.61$ .

- Determine SSE.
- Determine SST.
- Complete an ANOVA table.
- At the .05 significance level, is there a difference in the treatment means?
- If the mean for the hard rock station is 51.32 and the mean for the country/western station is 50.85, determine if there is a difference using the .05 significance level.

## DATA ANALYTICS

(The data for these exercises are available in Connect.)

33. **FILE** The North Valley Real Estate data report information on the homes sold last year.
- At the .02 significance level, is there a difference in the variability of the selling prices of the homes that have a pool versus those that do not have a pool?
  - At the .02 significance level, is there a difference in the variability of the selling prices of the homes with an attached garage versus those that do not have an attached garage?
  - At the .05 significance level, is there a difference in the mean selling price of the homes among the five townships?

- d. Adam Marty recently joined North Valley Real Estate and was assigned 20 homes to market and show. When he was hired, North Valley assured him that the 20 homes would be fairly assigned to him. When he reviewed the selling prices of his assigned homes, he thought that the prices were much below the average of \$357,000. Adam was able to find the data of the homes assigned to agents in the firm. Use statistical inference to compare the mean price of homes assigned to him to the mean price of homes assigned to the other agents. What do the results indicate? How is your analysis defining fairness?
34. **FILE** Refer to the Baseball 2016 data, which report information on the 30 Major League Baseball teams for the 2016 season.
- At the .10 significance level, is there a difference in the variation in team salary among the American and National League teams?
  - Create a variable that classifies a team's total attendance into three groups: less than 2.0 (million), 2.0 up to 3.0, and 3.0 or more. At the .05 significance level, is there a difference in the mean number of games won among the three groups?
  - Using the same attendance variable developed in part (b), is there a difference in the mean number of home runs hit per team? Use the .05 significance level.
  - Using the same attendance variable developed in part (b), is there a difference in the mean salary of the three groups? Use the .05 significance level.
35. **FILE** Refer to the Lincolnville School District bus data.
- Conduct a test of hypothesis to reveal whether the mean maintenance cost is equal for each of the bus manufacturers. Use the .01 significance level.
  - Conduct a test of hypothesis to determine whether the mean miles traveled since the last maintenance is equal for each bus manufacturer. Use the .05 significance level.

## PRACTICE TEST

### Part 1—Objective

- The test statistic for comparing two population variances follows the \_\_\_\_\_. (*F* distribution, *t* distribution, *z* distribution)
- The shape of the *F* distribution is \_\_\_\_\_. (symmetric, positively skewed, negatively skewed, uniform)
- The *F*-statistic is computed as the ratio of two \_\_\_\_\_.
- Analysis of variance (ANOVA) is used to compare two or more \_\_\_\_\_. (means, proportions, sample sizes, *z* values)
- The ANOVA test assumes equal \_\_\_\_\_. (population means, population standard deviations, sample sizes, *z* values)
- One-way ANOVA partitions total variation into two parts. One is called treatment variation, and the other is \_\_\_\_\_.
- In one-way ANOVA, the null hypothesis is that the population means are \_\_\_\_\_.
- A mean square is computed as a sum of squares divided by the \_\_\_\_\_.
- In one-way ANOVA, differences between treatment means are tested with \_\_\_\_\_. (confidence intervals, *z* values, significance levels, variances)
- For a one-way ANOVA, the treatments must be \_\_\_\_\_. (equal, independent, proportional, none of these)

### Part 2—Problems

- Is the variance of the distance traveled per week by two taxi cab companies operating in the Grand Strand area different? The *Sun News*, the local newspaper, is investigating and obtained the following sample information. Using the .10 significance level, is there a difference in the variance of the miles traveled?

| Variable           | Yellow Cab | Horse and Buggy Cab |
|--------------------|------------|---------------------|
| Mean miles         | 837        | 797                 |
| Standard deviation | 30         | 40                  |
| Sample size        | 14         | 12                  |

2. The results of a one-way ANOVA are reported below.

| ANOVA               |              |           |           |          |
|---------------------|--------------|-----------|-----------|----------|
| Source of Variation | <i>SS</i>    | <i>df</i> | <i>MS</i> | <i>F</i> |
| Between Groups      | 6.90         | 2         | 3.45      | 5.15     |
| Within Groups       | <u>12.04</u> | <u>18</u> | 0.67      |          |
| Total               | 18.94        | 20        |           |          |

- a. How many treatments are in the study?
- b. What is the total sample size?
- c. What is the critical value of  $F$ ?
- d. Write out the null hypothesis and the alternate hypothesis.
- e. What is your decision regarding the null hypothesis?
- f. Can we conclude any of the treatment means differ?

# Correlation and Linear Regression

# 13



©Ingram Publishing/SuperStock RF

- ▲ **TRAVELAIR.COM** samples domestic airline flights to explore the relationship between airfare and distance. The service would like to know if there is a correlation between airfare and flight distance. If there is a correlation, what percentage of the variation in airfare is accounted for by distance? How much does each additional mile add to the fare? (See Exercise 61 and **LO13-2**, **LO13-3**, and **LO13-5**.)

## LEARNING OBJECTIVES

---

*When you have completed this chapter, you will be able to:*

- LO13-1** Explain the purpose of correlation analysis.
- LO13-2** Calculate a correlation coefficient to test and interpret the relationship between two variables.
- LO13-3** Apply regression analysis to estimate the linear relationship between two variables.
- LO13-4** Evaluate the significance of the slope of the regression equation.
- LO13-5** Evaluate a regression equation's ability to predict using the standard estimate of the error and the coefficient of determination.
- LO13-6** Calculate and interpret confidence and prediction intervals.
- LO13-7** Use a log function to transform a nonlinear relationship.

## INTRODUCTION

Chapters 2 through 4 presented *descriptive statistics*. We organized raw data into a frequency distribution and computed several measures of location and measures of dispersion to describe the major characteristics of the distribution. In Chapters 5 through 7, we described probability, and from probability statements, we created probability distributions. In Chapters 8 through 12, we studied *statistical inference*, where we collected a sample to estimate a population parameter such as the population mean or population proportion. In addition, we used the sample data to test a hypothesis about a population mean or a population proportion, the difference between two population means, or the equality of several population means. Each of these tests involved just *one* interval- or ratio-level variable, such as the profit made on a car sale, the income of bank presidents, or the number of patients admitted each month to a particular hospital.

In this chapter, we shift the emphasis to the study of relationships between two interval- or ratio-level variables. In all business fields, identifying and studying relationships between variables can provide information on ways to increase profits, methods to decrease costs, or variables to predict demand. In marketing products, many firms use price reductions through coupons and discount pricing to increase sales. In this example, we are interested in the relationship between two variables: price reductions and sales. To collect the data, a company can test-market a variety of price reduction methods and observe sales. We hope to confirm a relationship that decreasing price leads to increased sales. In economics, you will find many relationships between two variables that are the basis of economics, such as price and demand.

As another familiar example, recall in Chapter 4 we used the Applewood Auto Group data to show the relationship between two variables with a scatter diagram. We plotted the profit for each vehicle sold on the vertical axis and the age of the buyer on the horizontal axis. See page 106. In that graph, we observed that as the age of the buyer increased, the profit for each vehicle also increased.

Other examples of relationships between two variables are:

- Does the amount Healthtex spends per month on training its sales force affect its monthly sales?
- Is the number of square feet in a home related to the cost to heat the home in January?
- In a study of fuel efficiency, is there a relationship between miles per gallon and the weight of a car?
- Does the number of hours that students study for an exam influence the exam score?

In this chapter, we carry this idea further. That is, we develop numerical measures to express the relationship between two variables. Is the relationship strong or weak? Is it direct or inverse? In addition, we develop an equation to express the relationship between variables. This will allow us to estimate one variable on the basis of another.

To begin our study of relationships between two variables, we examine the meaning and purpose of **correlation analysis**. We continue by developing an equation that will allow us to estimate the value of one variable based on the value of another. This is called **regression analysis**. We will also evaluate the ability of the equation to accurately make estimations.

### STATISTICS IN ACTION

The space shuttle *Challenger* exploded on January 28, 1986. An investigation of the cause examined four contractors: Rockwell International for the shuttle and engines, Lockheed Martin for ground support, Martin Marietta for the external fuel tanks, and Morton Thiokol for the solid fuel booster rockets. After several months, the investigation blamed the explosion on defective O-rings produced by Morton Thiokol. A study of the contractor's stock prices showed an interesting happenstance. On the day of the *Challenger* explosion, Morton Thiokol stock was down 11.86% and the stock of the other three lost only 2 to 3%. Can we conclude that financial markets predicted the outcome of the investigation?

### LO13-1

Explain the purpose of correlation analysis.

## WHAT IS CORRELATION ANALYSIS?

When we study the relationship between two interval- or ratio-scale variables, we often start with a scatter diagram. This procedure provides a visual representation of the relationship between the variables. The next step is usually to calculate the correlation coefficient. It provides a quantitative measure of the strength of the relationship between two variables. As an example, the sales manager of North American Copier Sales, which has a large sales force throughout the United States and Canada, wants to determine

**TABLE 13–1** Number of Sales Calls and Copiers Sold for 15 Salespeople

| Sales Representative | Sales Calls | Copiers Sold |
|----------------------|-------------|--------------|
| Brian Virost         | 96          | 41           |
| Carlos Ramirez       | 40          | 41           |
| Carol Saia           | 104         | 51           |
| Greg Fish            | 128         | 60           |
| Jeff Hall            | 164         | 61           |
| Mark Reynolds        | 76          | 29           |
| Meryl Rumsey         | 72          | 39           |
| Mike Kiel            | 80          | 50           |
| Ray Snarsky          | 36          | 28           |
| Rich Niles           | 84          | 43           |
| Ron Broderick        | 180         | 70           |
| Sal Spina            | 132         | 56           |
| Soni Jones           | 120         | 45           |
| Susan Welch          | 44          | 31           |
| Tom Keller           | 84          | 30           |

whether there is a relationship between the number of sales calls made in a month and the number of copiers sold that month. The manager selects a random sample of 15 representatives and determines, for each representative, the number of sales calls made and the number of copiers sold. This information is reported in Table 13–1.

By reviewing the data, we observe that there does seem to be some relationship between the number of sales calls and the number of units sold. That is, the salespeople who made the most sales calls sold the most units. However, the relationship is not “perfect” or exact. For example, Soni Jones made more sales calls than Carol Saia, but Carol sold more copiers.

In addition to the graphical techniques in Chapter 4, we will develop numerical measures to precisely describe the relationship between the two variables, sales calls and copiers sold. This group of statistical techniques is called **correlation analysis**.

**CORRELATION ANALYSIS** A group of techniques to measure the relationship between two variables.

The basic idea of correlation analysis is to report the relationship between two variables. The usual first step is to plot the data in a **scatter diagram**. An example will show how a scatter diagram is used.

### EXAMPLE

North American Copier Sales sells copiers to businesses of all sizes throughout the United States and Canada. Ms. Marcy Bancer was recently promoted to the position of national sales manager. At the upcoming sales meeting, the sales representatives from all over the country will be in attendance. She would like to impress upon them the importance of making that extra sales call each day. She decides to gather some information on the relationship between the number of sales calls and the number of copiers sold. She selects a random sample of 15 sales representatives and determines the number of sales calls they made last month and the number of copiers they sold. The sample information is reported in Table 13–1. What observations can you make about the relationship between the number of sales calls and the number of copiers sold? Develop a scatter diagram to display the information.



### SOLUTION

Based on the information in Table 13–1, Ms. Bancer suspects there is a relationship between the number of sales calls made in a month and the number of copiers sold. Ron Broderick sold the most copiers last month and made 180 sales calls. On the other hand, Ray Snarsky, Carlos Ramirez, and Susan Welch made the fewest calls: 36, 40, and 44. They also had among the lowest number of copiers sold of the sampled representatives.

The implication is that the number of copiers sold is related to the number of sales calls made. As the number of sales calls increases, it appears the number of copiers sold also increases. We refer to number of sales calls as the **independent variable** and number of copiers sold as the **dependent variable**.

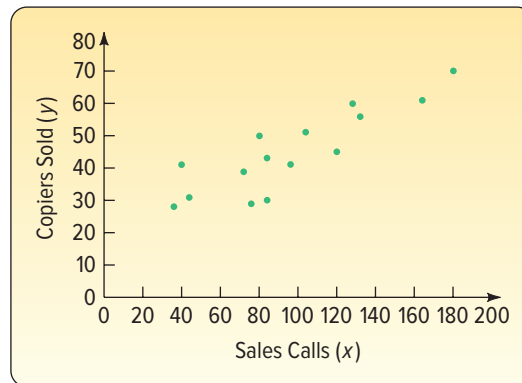
The independent variable provides the basis for estimating or predicting the dependent variable. For example, we would like to predict the expected number of copiers sold if a salesperson makes 100 sales calls. In the randomly selected sample data, the independent variable—sales calls—is a random number.

The dependent variable is the variable that is being predicted or estimated. It can also be described as the result or outcome for a particular value of the independent variable. The dependent variable is random. That is, for a given value of the independent variable, there are many possible outcomes for the dependent variable.

**INDEPENDENT VARIABLE** A variable that provides the basis for estimation.

**DEPENDENT VARIABLE** The variable that is being predicted or estimated.

It is common practice to scale the dependent variable (copiers sold) on the vertical or  $Y$ -axis and the independent variable (number of sales calls) on the horizontal or  $X$ -axis. To develop the scatter diagram of the North American Copier Sales information, we begin with the first sales representative, Brian Virost. Brian made 96 sales calls last month and sold 41 copiers, so  $x = 96$  and  $y = 41$ . To plot this point, move along the horizontal axis to  $x = 96$ , then go vertically to  $y = 41$  and place a dot at the intersection. This process is continued until all the paired data are plotted, as shown in Chart 13–1.



**CHART 13–1** Scatter Diagram Showing Sales Calls and Copiers Sold

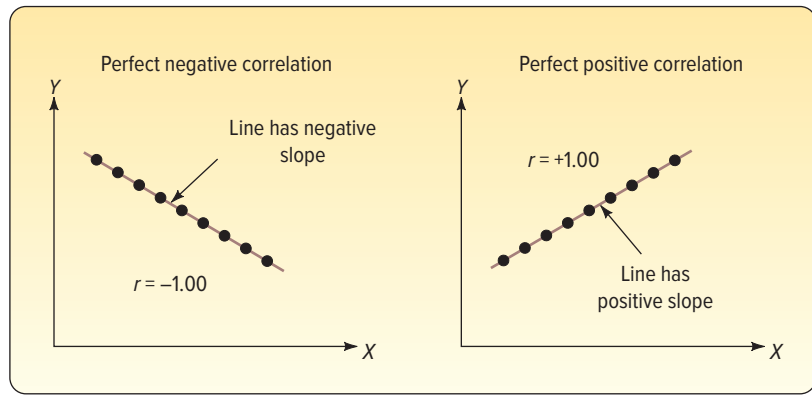
The scatter diagram shows graphically that the sales representatives who make more calls tend to sell more copiers. It is reasonable for Ms. Bancer, the national sales manager, to tell her salespeople that the more sales calls they make, the more copiers they can expect to sell. Note that, while there appears to be a positive relationship between the two variables, all the points do not fall on a straight line. In the following section, you will measure the strength and direction of this relationship between two variables by determining the correlation coefficient.

**LO13-2**

Calculate a correlation coefficient to test and interpret the relationship between two variables.

## THE CORRELATION COEFFICIENT

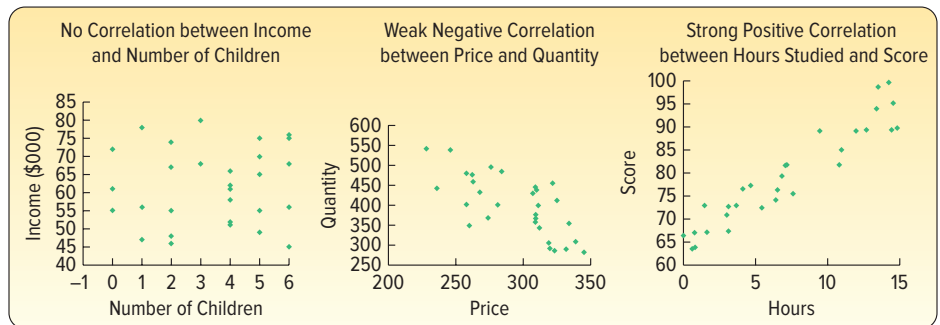
Originated by Karl Pearson about 1900, the **correlation coefficient** describes the strength of the relationship between two sets of interval-scaled or ratio-scaled variables. Designated  $r$ , it is often referred to as *Pearson's  $r$*  and as the *Pearson product-moment correlation coefficient*. It can assume any value from  $-1.00$  to  $+1.00$  inclusive. A correlation coefficient of  $-1.00$  or  $+1.00$  indicates *perfect correlation*. For example, a correlation coefficient for the preceding example computed to be  $+1.00$  would indicate that the number of sales calls and the number of copiers sold are perfectly related in a positive linear sense. A computed value of  $-1.00$  would reveal that sales calls and the number of copiers sold are perfectly related in an inverse linear sense. How the scatter diagram would appear if the relationship between the two variables were linear and perfect is shown in Chart 13–2.



**CHART 13–2** Scatter Diagrams Showing Perfect Negative Correlation and Perfect Positive Correlation

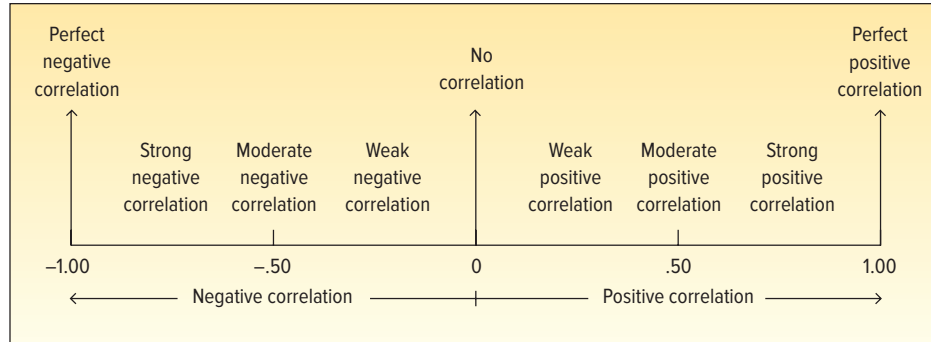
If there is absolutely no relationship between the two sets of variables, Pearson's  $r$  is zero. A correlation coefficient  $r$  close to 0 (say,  $.08$ ) shows that the linear relationship is quite weak. The same conclusion is drawn if  $r = -.08$ . Coefficients of  $-.91$  and  $+.91$  have equal strength; both indicate very strong correlation between the two variables. Thus, *the strength of the correlation does not depend on the direction (either  $-$  or  $+$ )*.

Scatter diagrams for  $r = 0$ , a weak  $r$  (say,  $-.23$ ), and a strong  $r$  (say,  $+.87$ ) are shown in Chart 13–3. Note that, if the correlation is weak, there is considerable scatter about a line drawn through the center of the data. For the scatter diagram representing a strong relationship, there is very little scatter about the line. This indicates, in the example shown on the chart, that hours studied is a good predictor of exam score.



**CHART 13–3** Scatter Diagrams Depicting Zero, Weak, and Strong Correlation

The following drawing summarizes the strength and direction of the correlation coefficient.



**CORRELATION COEFFICIENT** A measure of the strength of the linear relationship between two variables.

The characteristics of the correlation coefficient are summarized below.

**CHARACTERISTICS OF THE CORRELATION COEFFICIENT**

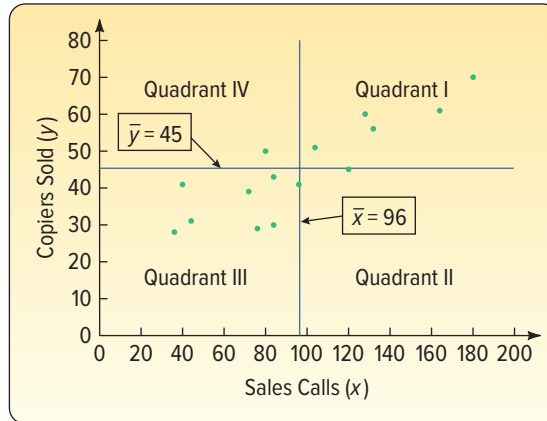
1. The sample correlation coefficient is identified by the lowercase letter *r*.
2. It shows the direction and strength of the linear relationship between two interval- or ratio-scale variables.
3. It ranges from  $-1$  up to and including  $+1$ .
4. A value near 0 indicates there is little linear relationship between the variables.
5. A value near 1 indicates a direct or positive linear relationship between the variables.
6. A value near  $-1$  indicates an inverse or negative linear relationship between the variables.

How is the value of the correlation coefficient determined? We will use the North American Copier Sales in Table 13–1 as an example. It is replicated in Table 13–2 for your convenience.

**TABLE 13–2** Number of Sales Calls and Copiers Sold for 15 Salespeople

| Sales Representative | Sales Calls | Copiers Sold |
|----------------------|-------------|--------------|
| Brian Virost         | 96          | 41           |
| Carlos Ramirez       | 40          | 41           |
| Carol Saia           | 104         | 51           |
| Greg Fish            | 128         | 60           |
| Jeff Hall            | 164         | 61           |
| Mark Reynolds        | 76          | 29           |
| Meryl Rumsey         | 72          | 39           |
| Mike Kiel            | 80          | 50           |
| Ray Snarsky          | 36          | 28           |
| Rich Niles           | 84          | 43           |
| Ron Broderick        | 180         | 70           |
| Sal Spina            | 132         | 56           |
| Soni Jones           | 120         | 45           |
| Susan Welch          | 44          | 31           |
| Tom Keller           | 84          | 30           |
| Total                | 1,440       | 675          |

We begin with a scatter diagram, similar to Chart 13–2. Draw a vertical line through the data values at the mean of the  $x$  values and a horizontal line at the mean of the  $y$  values. In Chart 13–4, we’ve added a vertical line at 96 calls ( $\bar{x} = \Sigma x/n = 1440/15 = 96$ ) and a horizontal line at 45 copiers ( $\bar{y} = \Sigma y/n = 675/15 = 45$ ). These lines pass through the “center” of the data and divide the scatter diagram into four quadrants. Think of moving the origin from  $(0, 0)$  to  $(96, 45)$ .



**CHART 13–4** Computation of the Correlation Coefficient

Two variables are positively related when the number of copiers sold is above the mean and the number of sales calls is also above the mean. These points appear in the upper-right quadrant (labeled Quadrant I) of Chart 13–4. Similarly, when the number of copiers sold is less than the mean, so is the number of sales calls. These points fall in the lower-left quadrant of Chart 13–4 (labeled Quadrant III). For example, the third person on the list in Table 13–2, Carol Saia, made 104 sales calls and sold 51 copiers. These values are above their respective means, so this point is located in Quadrant I, which is in the upper-right quadrant. She made 8 more calls than the mean number of sales calls and sold 6 more than the mean number sold. Tom Keller, the last name on the list in Table 13–2, made 84 sales calls and sold 30 copiers. Both of these values are less than their respective means, hence this point is in the lower-left quadrant. Tom made 12 fewer sales calls and sold 15 fewer copiers than the respective means. The deviations from the mean number of sales calls and the mean number of copiers sold are summarized in Table 13–3 for the

**TABLE 13–3** Deviations from the Mean and Their Products

| Sales Representative | Sales Calls ( $x$ ) | Copiers Sold ( $y$ ) | $x - \bar{x}$ | $y - \bar{y}$ | $(x - \bar{x})(y - \bar{y})$ |
|----------------------|---------------------|----------------------|---------------|---------------|------------------------------|
| Brian Virost         | 96                  | 41                   | 0             | -4            | 0                            |
| Carlos Ramirez       | 40                  | 41                   | -56           | -4            | 224                          |
| Carol Saia           | 104                 | 51                   | 8             | 6             | 48                           |
| Greg Fish            | 128                 | 60                   | 32            | 15            | 480                          |
| Jeff Hall            | 164                 | 61                   | 68            | 16            | 1,088                        |
| Mark Reynolds        | 76                  | 29                   | -20           | -16           | 320                          |
| Meryl Rumsey         | 72                  | 39                   | -24           | -6            | 144                          |
| Mike Kiel            | 80                  | 50                   | -16           | 5             | -80                          |
| Ray Snarsky          | 36                  | 28                   | -60           | -17           | 1,020                        |
| Rich Niles           | 84                  | 43                   | -12           | -2            | 24                           |
| Ron Broderick        | 180                 | 70                   | 84            | 25            | 2,100                        |
| Sal Spina            | 132                 | 56                   | 36            | 11            | 396                          |
| Soni Jones           | 120                 | 45                   | 24            | 0             | 0                            |
| Susan Welch          | 44                  | 31                   | -52           | -14           | 728                          |
| Tom Keller           | 84                  | 30                   | -12           | -15           | 180                          |
| Totals               | 1,440               | 675                  | 0             | 0             | 6,672                        |

15 sales representatives. The sum of the products of the deviations from the respective means is 6,672. That is, the term  $\Sigma(x - \bar{x})(y - \bar{y}) = 6,672$ .

In both the upper-right and the lower-left quadrants, the product of  $(x - \bar{x})(y - \bar{y})$  is positive because both of the factors have the same sign. In our example, this happens for all sales representatives except Mike Kiel. Mike made 80 sales calls (which is less than the mean) but sold 50 machines (which is more than the mean). We can therefore expect the correlation coefficient to have a positive value.

If the two variables are inversely related, one variable will be above the mean and the other below the mean. Most of the points in this case occur in the upper-left and lower-right quadrants, that is, Quadrants II and IV. Now  $(x - \bar{x})$  and  $(y - \bar{y})$  will have opposite signs, so their product is negative. The resulting correlation coefficient is negative.

What happens if there is no linear relationship between the two variables? The points in the scatter diagram will appear in all four quadrants. The negative products of  $(x - \bar{x})(y - \bar{y})$  offset the positive products, so the sum is near zero. This leads to a correlation coefficient near zero. So, the term  $\Sigma(x - \bar{x})(y - \bar{y})$  drives the strength as well as the sign of the relationship between the two variables.

The correlation coefficient is also unaffected by the units of the two variables. For example, if we had used hundreds of copiers sold instead of the number sold, the correlation coefficient would be the same. The correlation coefficient is independent of the scale used if we divide the term  $\Sigma(x - \bar{x})(y - \bar{y})$  by the sample standard deviations. It is also made independent of the sample size and bounded by the values +1.00 and -1.00 if we divide by  $(n - 1)$ .

This reasoning leads to the following formula:

**CORRELATION COEFFICIENT** **(13-1)**

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y}$$

To compute the correlation coefficient, we use the standard deviations of the sample of 15 sales calls and 15 copiers sold. We could use formula (3-8) to calculate the sample standard deviations, or we could use a statistical software package. For the specific Excel and Minitab commands, see the **Software Commands** in Appendix C. The following is the Excel output. The standard deviation of the number of sales calls is 42.76 and of the number of copiers sold is 12.89.

|    | A              | B                    | C               | D                | E                  | F                                | G      | H |
|----|----------------|----------------------|-----------------|------------------|--------------------|----------------------------------|--------|---|
| 1  |                | Sales Representative | Sales Calls (x) | Copiers Sold (y) |                    | Sales Calls (x) Copiers Sold (y) |        |   |
| 2  | Brian Virost   | 96                   | 41              |                  | Mean               | 96.00                            | 45.00  |   |
| 3  | Carlos Ramirez | 40                   | 41              |                  | Standard Error     | 11.04                            | 3.33   |   |
| 4  | Carol Saia     | 104                  | 51              |                  | Median             | 84.00                            | 43.00  |   |
| 5  | Greg Fish      | 128                  | 60              |                  | Mode               | 84.00                            | 41.00  |   |
| 6  | Jeff Hall      | 164                  | 61              |                  | Standard Deviation | 42.76                            | 12.89  |   |
| 7  | Mark Reynolds  | 76                   | 29              |                  | Sample Variance    | 1828.57                          | 166.14 |   |
| 8  | Meryl Rumsey   | 72                   | 39              |                  | Kurtosis           | -0.32                            | -0.73  |   |
| 9  | Mike Kiel      | 80                   | 50              |                  | Skewness           | 0.46                             | 0.36   |   |
| 10 | Ray Snarsky    | 36                   | 28              |                  | Range              | 144.00                           | 42.00  |   |
| 11 | Rich Niles     | 84                   | 43              |                  | Minimum            | 36.00                            | 28.00  |   |
| 12 | Ron Broderick  | 180                  | 70              |                  | Maximum            | 180.00                           | 70.00  |   |
| 13 | Sal Spina      | 132                  | 56              |                  | Sum                | 1440.00                          | 675.00 |   |
| 14 | Soni Jones     | 120                  | 45              |                  | Count              | 15.00                            | 15.00  |   |
| 15 | Susan Welch    | 44                   | 31              |                  |                    |                                  |        |   |
| 16 | Tom Keller     | 84                   | 30              |                  |                    |                                  |        |   |
| 17 | Total          | 1440                 | 675             |                  |                    |                                  |        |   |

Source: Microsoft Excel

We now insert these values into formula (13-1) to determine the correlation coefficient:

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y} = \frac{6672}{(15 - 1)(42.76)(12.89)} = 0.865$$

How do we interpret a correlation of 0.865? First, it is positive, so we conclude there is a direct relationship between the number of sales calls and the number of copiers sold. This confirms our reasoning based on the scatter diagram, Chart 13–4. The value of 0.865 is fairly close to 1.00, so we conclude that the association is strong.

We must be careful with the interpretation. The correlation of 0.865 indicates a strong positive linear association between the variables. Ms. Bancer would be correct to encourage the sales personnel to make that extra sales call because the number of sales calls made is related to the number of copiers sold. However, does this mean that more sales calls *cause* more sales? No, we have not demonstrated cause and effect here, only that the two variables—sales calls and copiers sold—are statistically related.

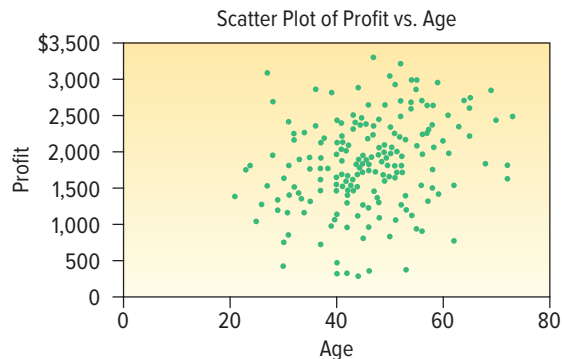
If there is a strong relationship (say, .97) between two variables, we are tempted to assume that an increase or decrease in one variable *causes* a change in the other variable. For example, historically, the consumption of Georgia peanuts and the consumption of aspirin have had a strong correlation. However, this does not indicate that an increase in the consumption of peanuts *caused* the consumption of aspirin to increase. Likewise, the incomes of professors and the number of inmates in mental institutions have increased proportionately. Further, as the population of donkeys has decreased, there has been an increase in the number of doctoral degrees granted. Relationships such as these are called **spurious correlations**. What we can conclude when we find two variables with a strong correlation is that there is a relationship or association between the two variables, not that a change in one causes a change in the other.

### EXAMPLE

The Applewood Auto Group’s marketing department believes younger buyers purchase vehicles on which lower profits are earned and older buyers purchase vehicles on which higher profits are earned. They would like to use this information as part of an upcoming advertising campaign to try to attract older buyers, for whom the profits tend to be higher. Develop a scatter diagram depicting the relationship between vehicle profits and age of the buyer. Use statistical software to determine the correlation coefficient. Would this be a useful advertising feature?

### SOLUTION

Using the Applewood Auto Group example, the first step is to graph the data using a scatter plot. It is shown in Chart 13–5.



**CHART 13–5** Scatter Diagram of Profit versus Age for the Applewood Auto Group Data

The scatter diagram suggests that a positive relationship does exist between age and profit; however, that relationship does not appear to be strong.

The next step is to calculate the correlation coefficient to evaluate the relative strength of the relationship. Statistical software provides an easy way to calculate the value of the correlation coefficient. The Excel output follows.

|   | A      | B          | C             |
|---|--------|------------|---------------|
| 1 |        | <b>Age</b> | <b>Profit</b> |
| 2 | Age    | 1          |               |
| 3 | Profit | 0.262      | 1             |

Source: Microsoft Excel

For these data,  $r = 0.262$ . To evaluate the relationship between a buyer’s age and the profit on a car sale:

1. The relationship is positive or direct. Why? Because the sign of the correlation coefficient is positive. This confirms that as the age of the buyer increases, the profit on a car sale also increases.
2. The correlation coefficient is:  $r = 0.262$ . It is much closer to zero than one. Therefore, the relationship between the two variables is weak. We would observe that the relationship between the age of a buyer and the profit of their purchase is not very strong.

For Applewood Auto Group, the data do not support a business decision to create an advertising campaign to attract older buyers.

### SELF-REVIEW 13-1



Haverty’s Furniture is a family business that has been selling to retail customers in the Chicago area for many years. The company advertises extensively on radio, TV, and the Internet, emphasizing low prices and easy credit terms. The owner would like to review the relationship between sales and the amount spent on advertising. Below is information on sales and advertising expense for the last four months.

| Month     | Advertising Expense (\$ million) | Sales Revenue (\$ million) |
|-----------|----------------------------------|----------------------------|
| July      | 2                                | 7                          |
| August    | 1                                | 3                          |
| September | 3                                | 8                          |
| October   | 4                                | 10                         |

- (a) The owner wants to forecast sales on the basis of advertising expense. Which variable is the dependent variable? Which variable is the independent variable?
- (b) Draw a scatter diagram.
- (c) Determine the correlation coefficient.
- (d) Interpret the strength of the correlation coefficient.

### EXERCISES

1. **FILE** The following sample of observations was randomly selected.

|          |   |   |   |   |    |
|----------|---|---|---|---|----|
| <b>x</b> | 4 | 5 | 3 | 6 | 10 |
| <b>y</b> | 4 | 6 | 5 | 7 | 7  |

Determine the correlation coefficient and interpret the relationship between  $x$  and  $y$ .

2. **FILE** The following sample of observations was randomly selected.

|     |    |    |   |    |    |    |   |   |
|-----|----|----|---|----|----|----|---|---|
| $x$ | 5  | 3  | 6 | 3  | 4  | 4  | 6 | 8 |
| $y$ | 13 | 15 | 7 | 12 | 13 | 11 | 9 | 5 |

Determine the correlation coefficient and interpret the relationship between  $x$  and  $y$ .

3. **FILE** Bi-lo Appliance Super-Store has outlets in several large metropolitan areas in New England. The general sales manager aired a commercial for a digital camera on selected local TV stations prior to a sale starting on Saturday and ending Sunday. She obtained the information for Saturday–Sunday digital camera sales at the various outlets and paired it with the number of times the advertisement was shown on the local TV stations. The purpose is to find whether there is any relationship between the number of times the advertisement was aired and digital camera sales. The pairings are:

| Location of TV Station | Number of Airings | Saturday–Sunday Sales (\$ thousands) |
|------------------------|-------------------|--------------------------------------|
| Providence             | 4                 | 15                                   |
| Springfield            | 2                 | 8                                    |
| New Haven              | 5                 | 21                                   |
| Boston                 | 6                 | 24                                   |
| Hartford               | 3                 | 17                                   |

- What is the dependent variable?
  - Draw a scatter diagram.
  - Determine the correlation coefficient.
  - Interpret these statistical measures.
4. **FILE** The production department of Celltronics International wants to explore the relationship between the number of employees who assemble a subassembly and the number produced. As an experiment, two employees were assigned to assemble the subassemblies. They produced 15 during a one-hour period. Then four employees assembled them. They produced 25 during a one-hour period. The complete set of paired observations follows.

| Number of Assemblers | One-Hour Production (units) |
|----------------------|-----------------------------|
| 2                    | 15                          |
| 4                    | 25                          |
| 1                    | 10                          |
| 5                    | 40                          |
| 3                    | 30                          |

The dependent variable is production; that is, it is assumed that different levels of production result from a different number of employees.

- Draw a scatter diagram.
  - Based on the scatter diagram, does there appear to be any relationship between the number of assemblers and production? Explain.
  - Compute the correlation coefficient.
5. **FILE** The city council of Pine Bluffs is considering increasing the number of police in an effort to reduce crime. Before making a final decision, the council asked the chief of police to survey other cities of similar size to determine the relationship between the number of police and the number of crimes reported. The chief gathered the following sample information.



| City        | Police | Number of Crimes | City      | Police | Number of Crimes |
|-------------|--------|------------------|-----------|--------|------------------|
| Oxford      | 15     | 17               | Holgate   | 17     | 7                |
| Starksville | 17     | 13               | Carey     | 12     | 21               |
| Danville    | 25     | 5                | Whistler  | 11     | 19               |
| Athens      | 27     | 7                | Woodville | 22     | 6                |

- Which variable is the dependent variable and which is the independent variable? Hint: Which of the following makes better sense: Cities with more police have fewer crimes, or cities with fewer crimes have more police? Explain your choice.
  - Draw a scatter diagram.
  - Determine the correlation coefficient.
  - Interpret the correlation coefficient. Does it surprise you that the correlation coefficient is negative?
6. **FILE** The owner of Maumee Ford-Volvo wants to study the relationship between the age of a car and its selling price. Listed below is a random sample of 12 used cars sold at the dealership during the last year.

| Car | Age (years) | Selling Price (\$000) | Car | Age (years) | Selling Price (\$000) |
|-----|-------------|-----------------------|-----|-------------|-----------------------|
| 1   | 9           | 8.1                   | 7   | 8           | 7.6                   |
| 2   | 7           | 6.0                   | 8   | 11          | 8.0                   |
| 3   | 11          | 3.6                   | 9   | 10          | 8.0                   |
| 4   | 12          | 4.0                   | 10  | 12          | 6.0                   |
| 5   | 8           | 5.0                   | 11  | 6           | 8.6                   |
| 6   | 7           | 10.0                  | 12  | 6           | 8.0                   |

- Draw a scatter diagram.
- Determine the correlation coefficient.
- Interpret the correlation coefficient. Does it surprise you that the correlation coefficient is negative?

## Testing the Significance of the Correlation Coefficient

Recall that the sales manager of North American Copier Sales found the correlation between the number of sales calls and the number of copiers sold was 0.865. This indicated a strong positive association between the two variables. However, only 15 salespeople were sampled. Could it be that the correlation in the population is actually 0? This would mean the correlation of 0.865 was due to chance, or sampling error. The population in this example is all the salespeople employed by the firm.

Resolving this dilemma requires a test to answer the question: Could there be zero correlation in the population from which the sample was selected? To put it another way, did the computed  $r$  come from a population of paired observations with zero correlation? To continue our convention of allowing Greek letters to represent a population parameter, we will let  $\rho$  represent the correlation in the population. It is pronounced "rho."

We will continue with the illustration involving sales calls and copiers sold. We employ the same hypothesis testing steps described in Chapter 10. The null hypothesis and the alternate hypothesis are:

$$\begin{aligned}
 H_0: \rho &= 0 && \text{(The correlation in the population is zero.)} \\
 H_1: \rho &\neq 0 && \text{(The correlation in the population is different from zero.)}
 \end{aligned}$$

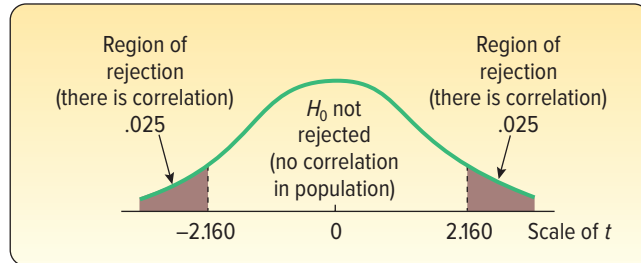
This is a two-tailed test. The null hypothesis can be rejected with either large or small sample values of the correlation coefficient.

The formula for  $t$  is:

**$t$  TEST FOR THE CORRELATION COEFFICIENT**

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \text{ with } n-2 \text{ degrees of freedom} \quad (13-2)$$

Using the .05 level of significance, the decision rule states that if the computed  $t$  falls in the area between plus 2.160 and minus 2.160, the null hypothesis is not rejected. To locate the critical value of 2.160, refer to Appendix B.5 for  $df = n - 2 = 15 - 2 = 13$ . See Chart 13-6.



**CHART 13-6** Decision Rule for Test of Hypothesis at .05 Significance Level and 13  $df$

Applying formula (13-2) to the example regarding the number of sales calls and units sold:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{.865\sqrt{15-2}}{\sqrt{1-.865^2}} = 6.216$$

The computed  $t$  is in the rejection region. Thus,  $H_0$  is rejected at the .05 significance level. Hence we conclude the correlation in the population is not zero. This indicates to the sales manager that there is correlation with respect to the number of sales calls made and the number of copiers sold in the population of salespeople.

We can also interpret the test of hypothesis in terms of  $p$ -values. A  $p$ -value is the likelihood of finding a value of the test statistic more extreme than the one computed, when  $H_0$  is true. To determine the  $p$ -value, go to the  $t$  distribution in Appendix B.5 and find the row for 13 degrees of freedom. The value of the test statistic is 6.216, so in the row for 13 degrees of freedom and a two-tailed test, find the value closest to 6.216. For a two-tailed test at the 0.001 significance level, the critical value is 4.221. Because 6.216 is greater than 4.221, we conclude that the  $p$ -value is less than 0.001.

Both Minitab and Excel will report the correlation between two variables. In addition to the correlation, Minitab reports the  $p$ -value for the test of hypothesis that the correlation in the population between the two variables is 0. The Minitab output follows.

|    | C1-T           | C2          | C3           | C4  | C5 | C6 | C7 | C8 | C9 | C10 | C11 |
|----|----------------|-------------|--------------|---|----|----|----|----|----|-----|-----|
|    | Name           | Sales Calls | Copiers Sold | Session   |    |    |    |    |    |     |     |
| 1  | Brian Virost   | 96          | 41           | Correlations: Sales Calls, Copiers Sold<br>Pearson correlation of Sales Calls and Copiers Sold = 0.865<br>P-Value = 0.000 |    |    |    |    |    |     |     |
| 2  | Carlos Ramirez | 40          | 41           |   |    |    |    |    |    |     |     |
| 3  | Carol Saia     | 104         | 51           |   |    |    |    |    |    |     |     |
| 4  | Greg Fish      | 128         | 60           |   |    |    |    |    |    |     |     |
| 5  | Jeff Hall      | 164         | 61           |   |    |    |    |    |    |     |     |
| 6  | Mark Reynolds  | 76          | 29           |   |    |    |    |    |    |     |     |
| 7  | Meryl Rumsey   | 72          | 39           |   |    |    |    |    |    |     |     |
| 8  | Mike Kiel      | 80          | 50           |   |    |    |    |    |    |     |     |
| 9  | Ray Snarsky    | 36          | 28           |   |    |    |    |    |    |     |     |
| 10 | Rich Niles     | 84          | 43           |   |    |    |    |    |    |     |     |
| 11 | Ron Broderick  | 180         | 70           |   |    |    |    |    |    |     |     |
| 12 | Sal Spina      | 132         | 56           |   |    |    |    |    |    |     |     |
| 13 | Soni Jones     | 120         | 45           |   |    |    |    |    |    |     |     |
| 14 | Susan Welch    | 44          | 31           |   |    |    |    |    |    |     |     |
| 15 | Tom Keller     | 84          | 30           |   |    |    |    |    |    |     |     |

Source: Minitab

### EXAMPLE

In the Applewood Auto Group example on page 373, we found that the correlation coefficient between the profit on the sale of a vehicle by the Applewood Auto Group and the age of the person that purchased the vehicle was 0.262. The sign of the correlation coefficient was positive, so we concluded there was a direct relationship between the two variables. However, because the value of the correlation coefficient was small—that is, near zero—we concluded that an advertising campaign directed toward the older buyers was not warranted. We can test our conclusion by conducting a hypothesis test that the correlation coefficient is greater than zero using the .05 significance level.

### SOLUTION

To test the hypothesis, we need to clarify the sample and population issues. Let's assume that the data collected on the 180 vehicles sold by the Applewood Auto Group is a sample from the population of *all* vehicles sold over many years by the Applewood Auto Group. The Greek letter  $\rho$  is the correlation coefficient in the population and  $r$  the correlation coefficient in the sample.

Our next step is to set up the null hypothesis and the alternate hypothesis. We test the null hypothesis that the correlation coefficient is equal to or less than zero. The alternate hypothesis is that there is positive correlation between the two variables.

$$\begin{aligned} H_0: \rho &\leq 0 && \text{(The correlation in the population is negative or equal to zero.)} \\ H_1: \rho &> 0 && \text{(The correlation in the population is positive.)} \end{aligned}$$

This is a one-tailed test because we are interested in confirming a positive association between the variables. The test statistic follows the  $t$  distribution with  $n - 2$  degrees of freedom, so the degrees of freedom are  $180 - 2 = 178$ . However, the value for 178 degrees of freedom is not in Appendix B.5. The closest value is 180, so we will use that value. Our decision rule is to reject the null hypothesis if the computed value of the test statistic is greater than 1.653.

We use formula (13-2) to find the value of the test statistic.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.262\sqrt{180-2}}{\sqrt{1-0.262^2}} = 3.622$$

Comparing the value of our test statistic of 3.622 to the critical value of 1.653, we reject the null hypothesis. We conclude that the sample correlation coefficient of 0.262 is too large to have come from a population with no correlation. To put our results another way, there is a positive correlation between profits and age in the population.

This result is confusing and seems contradictory. On one hand, we observed that the correlation coefficient did not indicate a very strong relationship and that the Applewood Auto Group marketing department should not use this information for its promotion and advertising decisions. On the other hand, the hypothesis test indicated that the correlation coefficient is not equal to zero and that a positive relationship between age and profit exists. How can this be? We must be very careful about the application of the hypothesis test results. The hypothesis test shows a statistically significant result. However, this result does not necessarily support a practical decision to start a new marketing and promotion campaign to older purchasers. In fact, the relatively low correlation coefficient is an indication that the outcome of a new marketing and promotion campaign to older potential purchasers is, at best, uncertain.

## SELF-REVIEW 13-2



A sample of 25 mayoral campaigns in medium-sized cities with populations between 50,000 and 250,000 showed that the correlation between the percent of the vote received and the amount spent on the campaign by the candidate was .43. At the .05 significance level, is there a positive association between the variables?

## EXERCISES

7. The following hypotheses are given.

$$H_0: \rho \leq 0$$

$$H_1: \rho > 0$$

A random sample of 12 paired observations indicated a correlation of .32. Can we conclude that the correlation in the population is greater than zero? Use the .05 significance level.

8. The following hypotheses are given.

$$H_0: \rho \geq 0$$

$$H_1: \rho < 0$$

A random sample of 15 paired observations has a correlation of  $-.46$ . Can we conclude that the correlation in the population is less than zero? Use the .05 significance level.

9. Pennsylvania Refining Company is studying the relationship between the pump price of gasoline and the number of gallons sold. For a sample of 20 stations last Tuesday, the correlation was .78. At the .01 significance level, is the correlation in the population greater than zero?
10. A study of 20 worldwide financial institutions showed the correlation between their assets and pretax profit to be .86. At the .05 significance level, can we conclude that there is positive correlation in the population?
11. The Airline Passenger Association studied the relationship between the number of passengers on a particular flight and the cost of the flight. It seems logical that more passengers on the flight will result in more weight and more luggage, which in turn will result in higher fuel costs. For a sample of 15 flights, the correlation between the number of passengers and total fuel cost was .667. Is it reasonable to conclude that there is positive association in the population between the two variables? Use the .01 significance level.
12. **FILE** The Student Government Association at Middle Carolina University wanted to demonstrate the relationship between the number of beers a student drinks and his or her blood alcohol content (BAC). A random sample of 18 students participated in a study in which each participating student was randomly assigned a number of 12-ounce cans of beer to drink. Thirty minutes after they consumed their assigned number of beers, a member of the local sheriff's office measured their blood alcohol content. The sample information is reported below.

| Student  | Beers | BAC  | Student  | Beers | BAC  |
|----------|-------|------|----------|-------|------|
| Charles  | 6     | 0.10 | Jaime    | 3     | 0.07 |
| Ellis    | 7     | 0.09 | Shannon  | 3     | 0.05 |
| Harriet  | 7     | 0.09 | Nellie   | 7     | 0.08 |
| Marlene  | 4     | 0.10 | Jeanne   | 1     | 0.04 |
| Tara     | 5     | 0.10 | Michele  | 4     | 0.07 |
| Kerry    | 3     | 0.07 | Seth     | 2     | 0.06 |
| Vera     | 3     | 0.10 | Gilberto | 7     | 0.12 |
| Pat      | 6     | 0.12 | Lillian  | 2     | 0.05 |
| Marjorie | 6     | 0.09 | Becky    | 1     | 0.02 |

Use a statistical software package to answer the following questions.

- Develop a scatter diagram for the number of beers consumed and BAC. Comment on the relationship. Does it appear to be strong or weak? Does it appear to be positive or inverse?
- Determine the correlation coefficient.
- At the .01 significance level, is it reasonable to conclude that there is a positive relationship in the population between the number of beers consumed and the BAC? What is the  $p$ -value?

### LO13-3

Apply regression analysis to estimate the linear relationship between two variables.

#### STATISTICS IN ACTION

In finance, investors are interested in the trade-off between returns and risk. One technique to quantify risk is a regression analysis of a company's stock price (dependent variable) and an average measure of the stock market (independent variable). Often the Standard and Poor's (S&P) 500 Index is used to estimate the market. The regression coefficient, called beta in finance, shows the change in a company's stock price for a one-unit change in the S&P Index. For example, if a stock has a beta of 1.5, then when the S&P index increases by 1%, the stock price will increase by 1.5%. The opposite is also true. If the S&P decreases by 1%, the stock price will decrease by 1.5%. If the beta is 1.0, then a 1% change in the index should show a 1% change in a stock price. If the beta is less than 1.0, then a 1% change in the index shows less than a 1% change in the stock price.

## REGRESSION ANALYSIS

In the previous sections of this chapter, we evaluated the direction and the significance of the linear relationship between two variables by finding the correlation coefficient. Regression analysis is another method to examine a linear relationship between two variables. This analysis uses the basic concepts of correlation but provides much more information by expressing the linear relationship between two variables in the form of an equation. Using this equation, we will be able to estimate the value of the dependent variable  $Y$  based on a selected value of the independent variable  $X$ . The technique used to develop the equation and provide the estimates is called **regression analysis**.



©Image Source/Getty Images RF

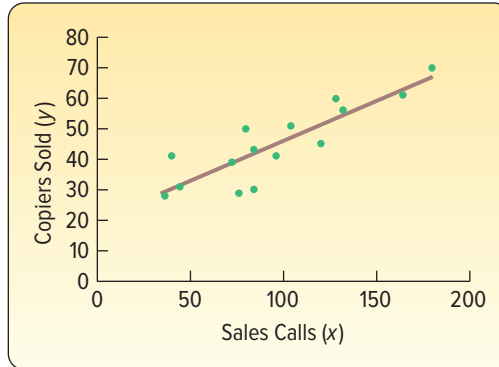
In Table 13–1, we reported the number of sales calls and the number of units sold for a sample of 15 sales representatives employed by North American Copier Sales. Chart 13–1 portrayed this information in a scatter diagram. Recall that we tested the significance of the correlation coefficient ( $r = 0.865$ ) and concluded that a significant relationship exists between the two variables. Now we want to develop a linear equation that expresses the relationship between the number of sales calls, the independent variable, and the number of units sold, the dependent variable. The equation for the line used to estimate  $Y$  on the basis of  $X$  is referred to as the **regression equation**.

**REGRESSION EQUATION** An equation that expresses the linear relationship between two variables.

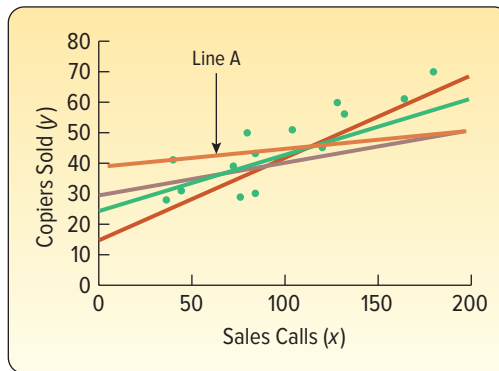
### Least Squares Principle

In regression analysis, our objective is to use the data to position a line that best represents the relationship between the two variables. Our first approach is to use a scatter diagram to visually position the line.

The scatter diagram in Chart 13–1 is reproduced in Chart 13–7, with a line drawn with a ruler through the dots to illustrate that a line would probably fit the data. However, the line drawn using a straight edge has one disadvantage: Its position is based in part on the judgment of the person drawing the line. The hand-drawn lines in Chart 13–8 represent the judgments of four people. All the lines except line  $A$  seem to be reasonable. That is, each line is centered among the graphed data. However, each would result in a different estimate of units sold for a particular number of sales calls.



**CHART 13-7** Sales Calls and Copiers Sold for 15 Sales Representatives



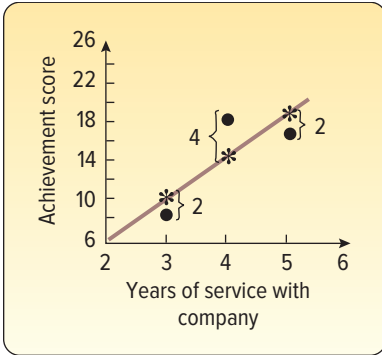
**CHART 13-8** Four Lines Superimposed on the Scatter Diagram

We would prefer a method that results in a single, best regression line. This method is called the **least squares principle**. It gives what is commonly referred to as the “best-fitting” line.

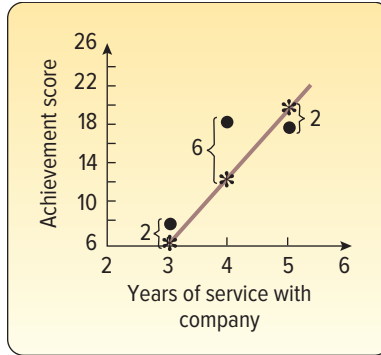
**LEAST SQUARES PRINCIPLE** A mathematical procedure that uses the data to position a line with the objective of minimizing the sum of the squares of the vertical distances between the actual  $y$  values and the predicted values of  $y$ .

To illustrate this concept, the same data are plotted in the three charts that follow. The dots are the actual values of  $y$ , and the asterisks are the predicted values of  $y$  for a given value of  $x$ . The regression line in Chart 13-9 was determined using the least squares method. It is the best-fitting line because the sum of the squares of the vertical deviations about it is at a minimum. The first plot ( $x = 3$ ,  $y = 8$ ) deviates by 2 from the line, found by  $10 - 8$ . The deviation squared is 4. The squared deviation for the plot  $x = 4$ ,  $y = 18$  is 16. The squared deviation for the plot  $x = 5$ ,  $y = 16$  is 4. The sum of the squared deviations is 24, found by  $4 + 16 + 4$ .

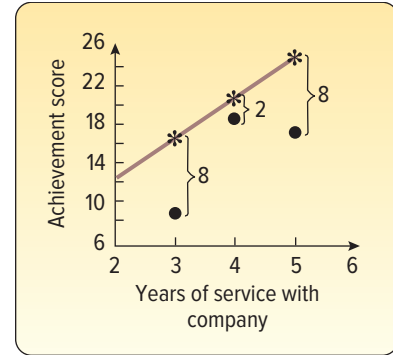
Assume that the lines in Charts 13-10 and 13-11 were drawn with a straight edge. The sum of the squared vertical deviations in Chart 13-10 is 44. For Chart 13-11, it is 132. Both sums are greater than the sum for the line in Chart 13-9, found by using the least squares method.



**CHART 13-9** The Least Squares Line



**CHART 13-10** Line Drawn with a Straight Edge



**CHART 13-11** Different Line Drawn with a Straight Edge

The equation of a line has the form

**GENERAL FORM OF LINEAR REGRESSION EQUATION**

$$\hat{y} = a + bx$$

**(13-3)**

where:

$\hat{y}$ , read *y* hat, is the estimated value of the *y* variable for a selected *x* value.

*a* is the *y*-intercept. It is the estimated value of *Y* when *x* = 0. Another way to put it is: *a* is the estimated value of *y* where the regression line crosses the *Y*-axis when *x* is zero.

*b* is the slope of the line, or the average change in  $\hat{y}$  for each change of one unit (either increase or decrease) in the independent variable *x*.

*x* is any value of the independent variable that is selected.

The general form of the linear regression equation is exactly the same form as the equation of any line. *a* is the *y*-intercept and *b* is the slope. The purpose of regression analysis is to calculate the values of *a* and *b* to develop a linear equation that best fits the data.

The formulas for *a* and *b* are:

**SLOPE OF THE REGRESSION LINE**

$$b = r \left( \frac{s_y}{s_x} \right)$$

**(13-4)**

where:

*r* is the correlation coefficient.

$s_y$  is the standard deviation of *y* (the dependent variable).

$s_x$  is the standard deviation of *x* (the independent variable).

**Y-INTERCEPT**

$$a = \bar{y} - b\bar{x}$$

**(13-5)**

where:

$\bar{y}$  is the mean of *y* (the dependent variable).

$\bar{x}$  is the mean of *x* (the independent variable).

**EXAMPLE**

Recall the example involving North American Copier Sales. The sales manager gathered information on the number of sales calls made and the number of copiers sold for a random sample of 15 sales representatives. As a part of her presentation at the upcoming sales meeting, Ms. Bancerc, the sales manager, would like to offer specific information about the relationship between the number of sales calls and the number of copiers sold. Use the least squares method to determine a linear equation to express the relationship between the two variables. What is the expected number of copiers sold by a representative who made 100 calls?

**SOLUTION**

The first step in determining the regression equation is to find the slope of the least squares regression line. That is, we need the value of  $b$ . In the previous section on page 372, we determined the correlation coefficient  $r$  (0.865). In the Excel output on page 372, we determined the standard deviation of the independent variable  $x$  (42.76) and the standard deviation of the dependent variable  $y$  (12.89). The values are inserted in formula (13–4).

$$b = r \left( \frac{s_y}{s_x} \right) = 0.865 \left( \frac{12.89}{42.76} \right) = 0.2608$$

Next, we need to find the value of  $a$ . To do this, we use the value for  $b$  that we just calculated as well as the means for the number of sales calls and the number of copiers sold. These means are also available in the Excel worksheet on page 372. From formula (13–5):

$$a = \bar{y} - b\bar{x} = 45 - 0.2608(96) = 19.9632$$

Thus, the regression equation is

$$\hat{y} = 19.9632 + 0.2608x.$$

So if a salesperson makes 100 calls, he or she can expect to sell 46.0432 copiers, found by

$$\hat{y} = 19.9632 + 0.2608x = 19.9632 + 0.2608(100) = 46.0432$$

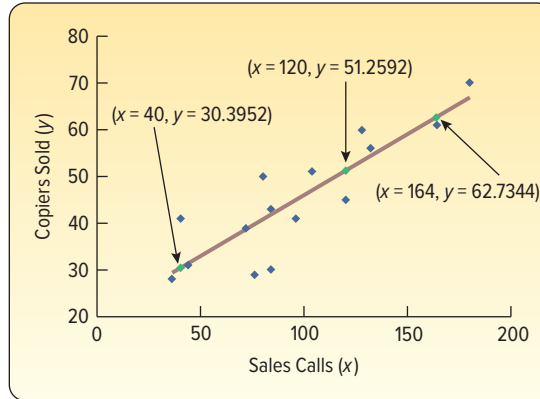
The  $b$  value of 0.2608 indicates that for each additional sales call, the sales representative can expect to increase the number of copiers sold by about 0.2608. To put it another way, 20 additional sales calls in a month will result in about five more copiers being sold, found by  $0.2608(20) = 5.216$ .

The  $a$  value of 19.9632 is the point where the equation crosses the  $Y$ -axis. A literal translation is that if no sales calls are made, that is  $x = 0$ , 19.9632 copiers will be sold. Note that  $x = 0$  is outside the range of values included in the sample and, therefore, should not be used to estimate the number of copiers sold. The sales calls ranged from 36 to 180, so estimates should be limited to that range.

**Drawing the Regression Line**

The least squares equation  $\hat{y} = 19.9632 + 0.2608x$  can be drawn on the scatter diagram. The fifth sales representative in the sample is Jeff Hall. He made 164 calls. His estimated number of copiers sold is  $\hat{y} = 19.9632 + 0.2608(164) = 62.7344$ . The plot  $x = 164$  and  $\hat{y} = 62.7344$  is located by moving to 164 on the  $X$ -axis and then going vertically to 62.7344. The other points on the regression equation can be determined





**CHART 13–12** The Line of Regression Drawn on the Scatter Diagram

by substituting a particular value of  $x$  into the regression equation and calculating  $\hat{y}$ . All the points are connected to give the line. See Chart 13–12.

| Sales Representative | Sales Calls ( $x$ ) | Copiers Sold ( $y$ ) | Estimated Sales ( $\hat{y}$ ) |
|----------------------|---------------------|----------------------|-------------------------------|
| Brian Virost         | 96                  | 41                   | 45.0000                       |
| Carlos Ramirez       | 40                  | 41                   | 30.3952                       |
| Carol Saia           | 104                 | 51                   | 47.0864                       |
| Greg Fish            | 128                 | 60                   | 53.3456                       |
| Jeff Hall            | 164                 | 61                   | 62.7344                       |
| Mark Reynolds        | 76                  | 29                   | 39.7840                       |
| Meryl Rumsey         | 72                  | 39                   | 38.7408                       |
| Mike Kiel            | 80                  | 50                   | 40.8272                       |
| Ray Snarsky          | 36                  | 28                   | 29.3520                       |
| Rich Niles           | 84                  | 43                   | 41.8704                       |
| Ron Broderick        | 180                 | 70                   | 66.9072                       |
| Sal Spina            | 132                 | 56                   | 54.3888                       |
| Soni Jones           | 120                 | 45                   | 51.2592                       |
| Susan Welch          | 44                  | 31                   | 31.4384                       |
| Tom Keller           | 84                  | 30                   | 41.8704                       |

The least squares regression line has some interesting and unique features. First, it will always pass through the point  $(\bar{x}, \bar{y})$ . To show this is true, we can use the mean number of sales calls to predict the number of copiers sold. In this example, the mean number of sales calls is 96, found by  $\bar{x} = 1,440/15$ . The mean number of copiers sold is 45.0, found by  $\bar{y} = 675/15$ . If we let  $x = 96$  and then use the regression equation to find the estimated value for  $\hat{y}$ , the result is:

$$\hat{y} = 19.9632 + 0.2608(96) = 45$$

The estimated number of copiers sold is exactly equal to the mean number of copiers sold. This example shows the regression line will pass through the point represented by the two means. In this case, the regression equation will pass through the point  $x = 96$  and  $y = 45$ .

Second, as we discussed earlier in this section, there is no other line through the data where the sum of the squared deviations is smaller. To put it another way, the term  $\sum(y - \hat{y})^2$  is smaller for the least squares regression equation than for any other equation. We use the Excel system to demonstrate this result in the following printout.

|    | A                | B                      | C                       | D                      | E              | F                          | G         | H                           | I          | J                             |
|----|------------------|------------------------|-------------------------|------------------------|----------------|----------------------------|-----------|-----------------------------|------------|-------------------------------|
| 1  | <b>Sales Rep</b> | <b>Sales Calls (x)</b> | <b>Copiers Sold (y)</b> | <b>Estimated Sales</b> | <b>(y - ŷ)</b> | <b>(y - ŷ)<sup>2</sup></b> | <b>y*</b> | <b>(y - y*)<sup>2</sup></b> | <b>y**</b> | <b>(y - y**) <sup>2</sup></b> |
| 2  | Brian Virost     | 96                     | 41                      | 45.0000                | -4.0000        | 16.0000                    | 44.4000   | 11.5600                     | 41.6000    | 0.3600                        |
| 3  | Carlos Ramirez   | 40                     | 41                      | 30.3952                | 10.6048        | 112.4618                   | 29.0000   | 144.0000                    | 29.0000    | 144.0000                      |
| 4  | Carol Saia       | 104                    | 51                      | 47.0864                | 3.9136         | 15.3163                    | 46.6000   | 19.3600                     | 43.4000    | 57.7600                       |
| 5  | Greg Fish        | 128                    | 60                      | 53.3456                | 6.6544         | 44.2810                    | 53.2000   | 46.2400                     | 48.8000    | 125.4400                      |
| 6  | Jeff Hall        | 164                    | 61                      | 62.7344                | -1.7344        | 3.0081                     | 63.1000   | 4.4100                      | 56.9000    | 16.8100                       |
| 7  | Mark Reynolds    | 76                     | 29                      | 39.7840                | -10.7840       | 116.2947                   | 38.9000   | 98.0100                     | 37.1000    | 65.6100                       |
| 8  | Meryl Rumsey     | 72                     | 39                      | 38.7408                | 0.2592         | 0.0672                     | 37.8000   | 1.4400                      | 36.2000    | 7.8400                        |
| 9  | Mike Kiel        | 80                     | 50                      | 40.8272                | 9.1728         | 84.1403                    | 40.0000   | 100.0000                    | 38.0000    | 144.0000                      |
| 10 | Ray Snarsky      | 36                     | 28                      | 29.3520                | -1.3520        | 1.8279                     | 27.9000   | 0.0100                      | 28.1000    | 0.0100                        |
| 11 | Rich Niles       | 84                     | 43                      | 41.8704                | 1.1296         | 1.2760                     | 41.1000   | 3.6100                      | 38.9000    | 16.8100                       |
| 12 | Ron Broderick    | 180                    | 70                      | 66.9072                | 3.0928         | 9.5654                     | 67.5000   | 6.2500                      | 60.5000    | 90.2500                       |
| 13 | Sal Spina        | 132                    | 56                      | 54.3888                | 1.6112         | 2.5960                     | 54.3000   | 2.8900                      | 49.7000    | 39.6900                       |
| 14 | Soni Jones       | 120                    | 45                      | 51.2592                | -6.2592        | 39.1776                    | 51.0000   | 36.0000                     | 47.0000    | 4.0000                        |
| 15 | Susan Welch      | 44                     | 31                      | 31.4384                | -0.4384        | 0.1922                     | 30.1000   | 0.8100                      | 29.9000    | 1.2100                        |
| 16 | Tom Keller       | 84                     | 30                      | 41.8704                | -11.8704       | 140.9064                   | 41.1000   | 123.2100                    | 38.9000    | 79.2100                       |
| 17 | <b>Total</b>     |                        |                         |                        | <b>0.0000</b>  | <b>587.1108</b>            |           | <b>597.8000</b>             |            | <b>793.0000</b>               |

Source: Microsoft Excel

In columns A, B, and C in the Excel spreadsheet above, we duplicated the sample information on sales and copiers sold from Table 13–1. In column D, we provide the estimated sales values, the  $\hat{y}$  values, as calculated above.

In column E, we calculate the **residuals**, or the error values. This is the difference between the actual values and the predicted values. That is, column E is  $(y - \hat{y})$ . For Soni Jones,

$$\hat{y} = 19.9632 + 0.2608(120) = 51.2592$$

Her actual value is 45. So the residual, or error of estimate, is

$$(y - \hat{y}) = (45 - 51.2592) = -6.2592$$

This value reflects the amount the predicted value of sales is “off” from the actual sales value.

Next, in column F, we square the residuals for each of the sales representatives and total the result. The total is 587.111.

$$\Sigma(y - \hat{y})^2 = 16.0000 + 112.4618 + \dots + 140.9064 = 587.1108$$

This is the sum of the squared differences or the least squares value. There is no other line through these 15 data points where the sum of the squared differences is smaller.

We can demonstrate the least squares criterion by choosing two arbitrary equations that are close to the least squares equation and determining the sum of the squared differences for these equations. In column G, we use the equation  $y^* = 18 + 0.275x$  to find the predicted value. Notice this equation is very similar to the least squares equation. In column H, we determine the residuals and square these residuals. For the first sales representative, Brian Virost,

$$y^* = 18 + 0.275(96) = 44.4$$

$$(y - y^*)^2 = (41 - 44.4)^2 = 11.56$$

This procedure is continued for the other 14 sales representatives and the squared residuals totaled. The result is 597.8. This is a larger value (597.8 is more than 587.1108) than the residuals for the least squares line.

In columns I and J on the output, we repeat the above process for yet another equation  $y^{**} = 20 + 0.225x$ . Again, this equation is similar to the least squares equation. The details for Brian Virost are:

$$y^{**} = 20 + 0.225x = 20 + 0.225(96) = 41.6$$

$$(y - y^{**})^2 = (41 - 41.6)^2 = 0.36$$

This procedure is continued for the other 14 sales representatives and the residuals totaled. The result is 793, which is also larger than the least squares values.

What have we shown with the example? The sum of the squared residuals  $[\sum(y - \hat{y})^2]$  for the least squares equation is smaller than for other selected lines. The bottom line is you will not be able to find a line passing through these data points where the sum of the squared residuals is smaller.

## SELF-REVIEW 13-3



Refer to Self-Review 13-1, where the owner of Haverty's Furniture Company was studying the relationship between sales and the amount spent on advertising. The advertising expense and sales revenue, both in millions of dollars, for the last 4 months are repeated below.

| Month     | Advertising Expense<br>(\$ million) | Sales Revenue<br>(\$ million) |
|-----------|-------------------------------------|-------------------------------|
| July      | 2                                   | 7                             |
| August    | 1                                   | 3                             |
| September | 3                                   | 8                             |
| October   | 4                                   | 10                            |

- Determine the regression equation.
- Interpret the values of  $a$  and  $b$ .
- Estimate sales when \$3 million is spent on advertising.

## EXERCISES

13. **FILE** The following sample of observations was randomly selected.

|     |   |   |   |   |    |
|-----|---|---|---|---|----|
| $x$ | 4 | 5 | 3 | 6 | 10 |
| $y$ | 4 | 6 | 5 | 7 | 7  |

- Determine the regression equation.
  - Determine the value of  $\hat{y}$  when  $x$  is 7.
14. **FILE** The following sample of observations was randomly selected.

|     |    |    |   |    |    |    |   |   |
|-----|----|----|---|----|----|----|---|---|
| $x$ | 5  | 3  | 6 | 3  | 4  | 4  | 6 | 8 |
| $y$ | 13 | 15 | 7 | 12 | 13 | 11 | 9 | 5 |

- Determine the regression equation.
  - Determine the value of  $\hat{y}$  when  $x$  is 7.
15. **FILE** Bradford Electric Illuminating Company is studying the relationship between kilowatt-hours (thousands) used and the number of rooms in a private single-family residence. A random sample of 10 homes yielded the following.

| Number of Rooms | Kilowatt-Hours (thousands) | Number of Rooms | Kilowatt-Hours (thousands) |
|-----------------|----------------------------|-----------------|----------------------------|
| 12              | 9                          | 8               | 6                          |
| 9               | 7                          | 10              | 8                          |
| 14              | 10                         | 10              | 10                         |
| 6               | 5                          | 5               | 4                          |
| 10              | 8                          | 7               | 7                          |

- Determine the regression equation.
- Determine the number of kilowatt-hours, in thousands, for a six-room house.

16. **FILE** Mr. James McWhinney, president of Daniel-James Financial Services, believes there is a relationship between the number of client contacts and the dollar amount of sales. To document this assertion, Mr. McWhinney gathered the following sample information. The  $x$  column indicates the number of client contacts last month and the  $y$  column shows the value of sales (\$ thousands) last month for each client sampled.

| Number of<br>Contacts,<br>$x$ | Sales<br>(\$ thousands),<br>$y$ | Number of<br>Contacts,<br>$x$ | Sales<br>(\$ thousands),<br>$y$ |
|-------------------------------|---------------------------------|-------------------------------|---------------------------------|
| 14                            | 24                              | 23                            | 30                              |
| 12                            | 14                              | 48                            | 90                              |
| 20                            | 28                              | 50                            | 85                              |
| 16                            | 30                              | 55                            | 120                             |
| 46                            | 80                              | 50                            | 110                             |

- a. Determine the regression equation.  
b. Determine the estimated sales if 40 contacts are made.
17. **FILE** A recent article in *Bloomberg Businessweek* listed the “Best Small Companies.” We are interested in the current results of the companies’ sales and earnings. A random sample of 12 companies was selected and the sales and earnings, in millions of dollars, are reported below.

| Company                   | Sales<br>(\$ millions) | Earnings<br>(\$ millions) | Company               | Sales<br>(\$ millions) | Earnings<br>(\$ millions) |
|---------------------------|------------------------|---------------------------|-----------------------|------------------------|---------------------------|
| Papa John’s International | \$89.2                 | \$4.9                     | Checkmate Electronics | \$17.5                 | \$ 2.6                    |
| Applied Innovation        | 18.6                   | 4.4                       | Royal Grip            | 11.9                   | 1.7                       |
| IntegraCare               | 18.2                   | 1.3                       | M-Wave                | 19.6                   | 3.5                       |
| Wall Data                 | 71.7                   | 8.0                       | Serving-N-Slide       | 51.2                   | 8.2                       |
| Davidson & Associates     | 58.6                   | 6.6                       | Daig                  | 28.6                   | 6.0                       |
| Chico’s FAS               | 46.8                   | 4.1                       | Cobra Golf            | 69.2                   | 12.8                      |

Let sales be the independent variable and earnings be the dependent variable.

- a. Draw a scatter diagram.  
b. Compute the correlation coefficient.  
c. Determine the regression equation.  
d. For a small company with \$50.0 million in sales, estimate the earnings.
18. **FILE** We are studying mutual bond funds for the purpose of investing in several funds. For this particular study, we want to focus on the assets of a fund and its five-year performance. The question is: Can the five-year rate of return be estimated based on the assets of the fund? Nine mutual funds were selected at random, and their assets and rates of return are shown below.

| Fund                         | Assets<br>(\$ millions) | Return<br>(%) | Fund                     | Assets<br>(\$ millions) | Return<br>(%) |
|------------------------------|-------------------------|---------------|--------------------------|-------------------------|---------------|
| AARP High Quality Bond       | \$622.2                 | 10.8          | MFS Bond A               | \$494.5                 | 11.6          |
| Babson Bond L                | 160.4                   | 11.3          | Nichols Income           | 158.3                   | 9.5           |
| Compass Capital Fixed Income | 275.7                   | 11.4          | T. Rowe Price Short-term | 681.0                   | 8.2           |
| Galaxy Bond Retail           | 433.2                   | 9.1           | Thompson Income B        | 241.3                   | 6.8           |
| Keystone Custodian B-1       | 437.9                   | 9.2           |                          |                         |               |

- a. Draw a scatter diagram.
  - b. Compute the correlation coefficient.
  - c. Write a brief report of your findings for parts (a) and (b).
  - d. Determine the regression equation. Use assets as the independent variable.
  - e. For a fund with \$400.0 million in sales, determine the five-year rate of return (in percent).
19. **FILE** Refer to Exercise 5. Assume the dependent variable is number of crimes.
    - a. Determine the regression equation.
    - b. Estimate the number of crimes for a city with 20 police officers.
    - c. Interpret the regression equation.
  20. **FILE** Refer to Exercise 6.
    - a. Determine the regression equation.
    - b. Estimate the selling price of a 10-year-old car.
    - c. Interpret the regression equation.

**LO13-4**

Evaluate the significance of the slope of the regression equation.

## TESTING THE SIGNIFICANCE OF THE SLOPE

In the prior section, we showed how to find the equation of the regression line that best fits the data. The method for finding the equation is based on the *least squares principle*. The purpose of the regression equation is to quantify a linear relationship between two variables.

The next step is to analyze the regression equation by conducting a test of hypothesis to see if the slope of the regression line is different from zero. Why is this important? If we can show that the slope of the line in the population is different from zero, then we can conclude that using the regression equation adds to our ability to predict or forecast the dependent variable based on the independent variable. If we cannot demonstrate that this slope is different from zero, then we conclude there is no merit to using the independent variable as a predictor. To put it another way, if we cannot show the slope of the line is different from zero, we might as well use the mean of the dependent variable as a predictor, rather than use the regression equation.

Following from the hypothesis-testing procedure in Chapter 10, the null and alternative hypotheses are:

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

We use  $\beta$  (the Greek letter beta) to represent the population slope for the regression equation. This is consistent with our policy to identify population parameters by Greek letters. We assumed the information regarding North American Copier Sales, Table 13–2, is a sample. Be careful here. Remember, this is a single sample of salespeople, but when we selected a particular salesperson we identified two pieces of information: how many customers they called on and how many copiers they sold.

We identified the slope value as  $b$ . So  $b$  is our computed slope based on a sample and is an estimate of the population's slope, identified as  $\beta$ . The null hypothesis is that the slope of the regression equation in the population is zero. If this is the case, the regression line is horizontal and there is no relationship between the independent variable,  $X$ , and the dependent variable,  $Y$ . In other words, the value of the dependent variable is the same for any value of the independent variable and does not offer us any help in estimating the value of the dependent variable.

What if the null hypothesis is rejected? If the null hypothesis is rejected and the alternate hypothesis accepted, this indicates that the slope of the regression line for the population is not equal to zero. To put it another way, a significant relationship exists between the two variables. Knowing the value of the independent variable allows us to estimate the value of the dependent variable.

Before we test the hypothesis, we use statistical software to determine the needed regression statistics. We continue to use the North American Copier Sales data from

Table 13–2 and use Excel to perform the necessary calculations. The following spreadsheet shows three tables to the right of the sample data.

|    | A                    | B               | C                | D | E                     | F  | G           | H       | I       | J              |
|----|----------------------|-----------------|------------------|---|-----------------------|--|-------------|---------|---------|----------------|
| 1  | Sales Representative | Sales calls (x) | Copiers Sold (y) |   | SUMMARY OUTPUT        |  |             |         |         |                |
| 2  | Brian Virost         | 96              | 41               |   |                       |  |             |         |         |                |
| 3  | Carlos Ramirez       | 40              | 41               |   | Regression Statistics |  |             |         |         |                |
| 4  | Carol Saia           | 104             | 51               |   | Multiple R            | 0.865                                      |             |         |         |                |
| 5  | Greg Fish            | 128             | 60               |   | R Square              | 0.748                                      |             |         |         |                |
| 6  | Jeff Hall            | 164             | 61               |   | Adjusted R Square     | 0.728                                      |             |         |         |                |
| 7  | Mark Reynolds        | 76              | 29               |   | Standard Error        | 6.720                                      |             |         |         |                |
| 8  | Meryl Rumsey         | 72              | 39               |   | Observations          | 15   |             |         |         |                |
| 9  | Mike Kiel            | 80              | 50               |   |                       |  |             |         |         |                |
| 10 | Ray Snarsky          | 36              | 28               |   | ANOVA                 |  |             |         |         |                |
| 11 | Rich Niles           | 84              | 43               |   |                       | df   | SS          | MS      | F       | Significance F |
| 12 | Ron Broderick        | 180             | 70               |   | Regression            | 1  | 1738.89     | 1738.89 | 38.5031 | 3.19277E-05    |
| 13 | Sal Spina            | 132             | 56               |   | Residual              | 13   | 587.11      | 45.1623 |         |                |
| 14 | Sani Jones           | 120             | 45               |   | Total                 | 14   | 2326        |         |         |                |
| 15 | Susan Welch          | 44              | 31               |   |                       |  |             |         |         |                |
| 16 | Tom Keller           | 84              | 30               |   |                       | Coefficients Standard Error t Stat P-value |             |         |         |                |
| 17 |                      |                 |                  |   | Intercept             | 19.9800                                    | 4.389675533 | 4.55159 | 0.00054 |                |
| 18 |                      |                 |                  |   | Sales calls (x)       | 0.2606                                     | 0.042001817 | 6.20509 | 3.2E-05 |                |

Source: Microsoft Excel

- Starting on the top are the *Regression Statistics*. We will use this information later in the chapter, but notice that the “Multiple R” value is familiar. It is 0.865, which is the correlation coefficient we calculated using formula (13–1).
- Next is an ANOVA table. This is a useful table for summarizing regression information. We will refer to it later in this chapter and use it extensively in the next chapter when we study multiple regression.
- At the bottom, highlighted in blue, is the information needed to conduct our test of hypothesis regarding the slope of the line. It includes the value of the slope, which is 0.2606, and the intercept, which is 19.98. (Note that these values for the slope and the intercept are slightly different from those computed in the Example/Solution on page 383. These small differences are due to rounding.) In the column to the right of the regression coefficient is a column labeled “Standard Error.” This is a value similar to the standard error of the mean. Recall that the standard error of the mean reports the variation in the sample means. In a similar fashion, these standard errors report the possible variation in slope and intercept values. The standard error of the slope coefficient is 0.0420.

To test the null hypothesis, we use the  $t$ -distribution with  $(n - 2)$  and the following formula.

$$\text{TEST FOR THE SLOPE} \quad t = \frac{b - 0}{s_b} \quad \text{with } n - 2 \text{ degrees of freedom} \quad (13-6)$$

where:

$b$  is the estimate of the regression line’s slope calculated from the sample information.

$s_b$  is the standard error of the slope estimate, also determined from sample information.

Our first step is to set the null and the alternative hypotheses. They are:

$$H_0 : \beta \leq 0$$

$$H_1 : \beta > 0$$

Notice that we have a one-tailed test. If we do not reject the null hypothesis, we conclude that the slope of the regression line in the population could be zero. This means the independent variable is of no value in improving our estimate of the dependent

variable. In our case, this means that knowing the number of sales calls made by a representative does not help us predict the sales.

If we reject the null hypothesis and accept the alternative, we conclude the slope of the line is greater than zero. Hence, the independent variable is an aid in predicting the dependent variable. Thus, if we know the number of sales calls made by a salesperson, we can predict or forecast their sales. We also know, because we have demonstrated that the slope of the line is greater than zero—that is, positive—that more sales calls will result in the sale of more copiers.

The  $t$  distribution is the test statistic; there are 13 degrees of freedom, found by  $n - 2 = 15 - 2$ . We use the .05 significance level. From Appendix B.5, the critical value is 1.771. Our decision rule is to reject the null hypothesis if the value computed from formula (13–6) is greater than 1.771. We apply formula (13–6) to find  $t$ .

$$t = \frac{b - 0}{s_b} = \frac{0.2606 - 0}{0.042} = 6.205$$

The computed value of 6.205 exceeds our critical value of 1.771, so we reject the null hypothesis and accept the alternative hypothesis. We conclude that the slope of the line is greater than zero. The independent variable, number of sales calls, is useful in estimating copier sales.

The table also provides us information on the  $p$ -value of this test. This cell is highlighted in purple. So we could select a significance level, say .05, and compare that value with the  $p$ -value. In this case, the calculated  $p$ -value in the table is reported in exponential notation and is equal to 0.0000319, so our decision is to reject the null hypothesis. An important caution is that the  $p$ -values reported in the statistical software are usually for a *two-tailed test*.

Before moving on, here is an interesting note. When we conduct a hypothesis test regarding the correlation coefficient for these same data using formula (13–2) with the computed and not rounded correlation coefficient of 0.86463, we obtain the same value of the  $t$ -statistic,  $t = 6.205$ . Actually, when comparing the results of simple linear regression and correlation analysis, the two tests are equivalent and will always yield exactly the same values of  $t$  and the same  $p$ -values. Interesting!

## SELF-REVIEW 13-4



Refer to Self-Review 13–1, where the owner of Haverty's Furniture Company studied the relationship between the amount spent on advertising in a month and sales revenue for that month. The amount of sales is the dependent variable and advertising expense is the independent variable. The regression equation in that study was  $\hat{y} = 1.5 + 2.2x$  for a sample of 5 months. Conduct a test of hypothesis to show there is a positive relationship between advertising and sales. From statistical software, the standard error of the regression coefficient is 0.42. Use the .05 significance level.

## EXERCISES

21. **FILE** Refer to Exercise 5. The regression equation is  $\hat{y} = 29.29 - 0.96x$ , the sample size is 8, and the standard error of the slope is 0.22. Use the .05 significance level. Can we conclude that the slope of the regression line is less than zero?
22. **FILE** Refer to Exercise 6. The regression equation is  $\hat{y} = 11.18 - 0.49x$ , the sample size is 12, and the standard error of the slope is 0.23. Use the .05 significance level. Can we conclude that the slope of the regression line is less than zero?
23. **FILE** Refer to Exercise 17. The regression equation is  $\hat{y} = 1.85 + .08x$ , the sample size is 12, and the standard error of the slope is 0.03. Use the .05 significance level. Can we conclude that the slope of the regression line is *different from zero*?
24. **FILE** Refer to Exercise 18. The regression equation is  $\hat{y} = 9.9198 - 0.00039x$ , the sample size is 9, and the standard error of the slope is 0.0032. Use the .05 significance level. Can we conclude that the slope of the regression line is less than zero?

**LO13-5**

Evaluate a regression equation's ability to predict using the standard estimate of the error and the coefficient of determination.

## EVALUATING A REGRESSION EQUATION'S ABILITY TO PREDICT

### The Standard Error of Estimate

The results of the regression analysis for North American Copier Sales show a significant relationship between number of sales calls and the number of sales made. By substituting the names of the variables into the equation, it can be written as:

$$\text{Number of copiers sold} = 19.9632 + 0.2608 (\text{Number of sales calls})$$

The equation can be used to estimate the number of copiers sold for any given "number of sales calls" within the range of the data. For example, if the number of sales calls is 84, then we can predict the number of copiers sold. It is 41.8704, found by  $19.9632 + 0.2608(84)$ . However, the data show two sales representatives with 84 sales calls and 30 and 43 copiers sold. So, is the regression equation a good predictor of "number of copiers sold"?

Perfect prediction, which is finding the exact outcome, is practically impossible in almost all disciplines including economics and business. For example:

- A large electronics firm, with production facilities throughout the United States, has a stock option plan for employees. Suppose there is a relationship between the number of years employed and the number of shares owned. This relationship is likely because as number of years of service increases, the number of shares an employee earns also increases. If we observe all employees with 20 years of service, they would most likely own different numbers of shares.
- A real estate developer in the southwest United States studied the relationship between the income of buyers and the size, in square feet, of the home they purchased. The developer's analysis shows that as the income of a purchaser increases, the size of the home purchased will also increase. However, all buyers with an income of \$70,000 will not purchase a home of exactly the same size.

What is needed, then, is a measure that describes how precise the prediction of  $Y$  is based on  $X$  or, conversely, how inaccurate the estimate might be. This measure is called the **standard error of estimate**. The standard error of estimate is symbolized by  $s_{y \cdot x}$ . The subscript,  $y \cdot x$ , is interpreted as the standard error of  $y$  for a given value of  $x$ . It is the same concept as the standard deviation discussed in Chapter 3. The standard deviation measures the dispersion around the mean. The standard error of estimate measures the dispersion about the regression line for a given value of  $x$ .

**STANDARD ERROR OF ESTIMATE** A measure of the dispersion, or scatter, of the observed values around the line of regression for a given value of  $x$ .

The standard error of estimate is found using formula (13-7).

**STANDARD ERROR OF ESTIMATE**

$$s_{y \cdot x} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}} \quad (13-7)$$

The calculation of the standard error of estimate requires the sum of the squared differences between each observed value of  $y$  and the predicted value of  $y$ , which is identified as  $\hat{y}$  in the numerator. This calculation is illustrated in the following spreadsheet. See the highlighted cell in the lower-right corner.



|    | A                           | B                      | C                       | D                      | E              | F                          |
|----|-----------------------------|------------------------|-------------------------|------------------------|----------------|----------------------------|
| 1  | <b>Sales Representative</b> | <b>Sales Calls (x)</b> | <b>Copiers Sold (y)</b> | <b>Estimated Sales</b> | <b>(y - ŷ)</b> | <b>(y - ŷ)<sup>2</sup></b> |
| 2  | Brian Virost                | 96                     | 41                      | 45.0000                | -4.0000        | 16.0000                    |
| 3  | Carlos Ramirez              | 40                     | 41                      | 30.3952                | 10.6048        | 112.4618                   |
| 4  | Carol Saia                  | 104                    | 51                      | 47.0864                | 3.9136         | 15.3163                    |
| 5  | Greg Fish                   | 128                    | 60                      | 53.3456                | 6.6544         | 44.2810                    |
| 6  | Jeff Hall                   | 164                    | 61                      | 62.7344                | -1.7344        | 3.0081                     |
| 7  | Mark Reynolds               | 76                     | 29                      | 39.7840                | -10.7840       | 116.2947                   |
| 8  | Meryl Rumsey                | 72                     | 39                      | 38.7408                | 0.2592         | 0.0672                     |
| 9  | Mike Kiel                   | 80                     | 50                      | 40.8272                | 9.1728         | 84.1403                    |
| 10 | Ray Snarsky                 | 36                     | 28                      | 29.3520                | -1.3520        | 1.8279                     |
| 11 | Rich Niles                  | 84                     | 43                      | 41.8704                | 1.1296         | 1.2760                     |
| 12 | Ron Broderick               | 180                    | 70                      | 66.9072                | 3.0928         | 9.5654                     |
| 13 | Sal Spina                   | 132                    | 56                      | 54.3888                | 1.6112         | 2.5960                     |
| 14 | Soni Jones                  | 120                    | 45                      | 51.2592                | -6.2592        | 39.1776                    |
| 15 | Susan Welch                 | 44                     | 31                      | 31.4384                | -0.4384        | 0.1922                     |
| 16 | Tom Keller                  | 84                     | 30                      | 41.8704                | -11.8704       | 140.9064                   |
| 17 | <b>Total</b>                |                        |                         |                        | <b>0.0000</b>  | <b>587.1108</b>            |

Source: Microsoft Excel

The calculation of the standard error of estimate is:

$$s_{y \cdot x} = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}} = \sqrt{\frac{587.1108}{15 - 2}} = 6.720$$

The standard error of estimate can be calculated using statistical software such as Excel. It is included in Excel's regression analysis on page 389 and highlighted in yellow. Its value is 6.720.

If the standard error of estimate is small, this indicates that the data are relatively close to the regression line and the regression equation can be used to predict  $y$  with little error. If the standard error of estimate is large, this indicates that the data are widely scattered around the regression line and the regression equation will not provide a precise estimate of  $y$ .

## The Coefficient of Determination

Using the standard error of estimate provides a relative measure of a regression equation's ability to predict. We will use it to provide more specific information about a prediction in the next section. Another statistic, called the **coefficient of determination**, is easier to interpret and provides useful information on the regression equation's ability to predict.

**COEFFICIENT OF DETERMINATION** The proportion of the total variation in the dependent variable  $Y$  that is explained, or accounted for, by the variation in the independent variable  $X$ .

The coefficient of determination is easy to compute. It is the correlation coefficient squared. Therefore, the term  $R$ -square is also used. With the North American Copier Sales data, the correlation coefficient for the relationship between the number of copiers sold and the number of sales calls is 0.865. If we compute  $(0.865)^2$ , the coefficient of determination is 0.748. See the blue (Multiple R) and green ( $R$ -square) highlighted cells in the spreadsheet on page 389. To better interpret the coefficient of determination, convert it to a percentage. Hence, we say that 74.8% of the variation in the number of copiers sold is explained, or accounted for, by the variation in the number of sales calls.

How well can the regression equation predict number of copiers sold with number of sales calls made? If it were possible to make perfect predictions, the coefficient of determination would be 100%. That would mean that the independent variable, number of sales calls, explains or accounts for all the variation in the number of copiers sold. A coefficient of determination of 100% is associated with a correlation coefficient of +1.0 or -1.0. Refer to Chart 13-2, which shows that a perfect prediction is associated with a perfect linear relationship where all the data points form a perfect line in a scatter diagram. Our analysis shows that only 74.8% of the variation in copiers sold is explained by the number of sales calls. Clearly, these data do not form a perfect line. Instead, the data are scattered around the best-fitting, least squares regression line, and there will be error in the predictions. In the next section, the standard error of estimate is used to provide more specific information regarding the error associated with using the regression equation to make predictions.

## SELF-REVIEW 13-5



Refer to Self-Review 13-1, where the owner of Haverty's Furniture Company studied the relationship between the amount spent on advertising in a month and sales revenue for that month. The amount of sales is the dependent variable and advertising expense is the independent variable.

- Determine the standard error of estimate.
- Determine the coefficient of determination.
- Interpret the coefficient of determination.

## EXERCISES

(You may wish to use a statistical software package such as Excel, Minitab, or MegaStat to assist in your calculations.)

- Refer to Exercise 5. Determine the standard error of estimate and the coefficient of determination. Interpret the coefficient of determination.
- Refer to Exercise 6. Determine the standard error of estimate and the coefficient of determination. Interpret the coefficient of determination.
- Refer to Exercise 15. Determine the standard error of estimate and the coefficient of determination. Interpret the coefficient of determination.
- Refer to Exercise 16. Determine the standard error of estimate and the coefficient of determination. Interpret the coefficient of determination.

## Relationships among the Correlation Coefficient, the Coefficient of Determination, and the Standard Error of Estimate

In formula (13-7) shown on page 391, we described the standard error of estimate. Recall that it measures how close the actual values are to the regression line. When the standard error is small, it indicates that the two variables are closely related. In the calculation of the standard error, the key term is

$$\Sigma(y - \hat{y})^2$$

If the value of this term is small, then the standard error will also be small.

The correlation coefficient measures the strength of the linear association between two variables. When the points on the scatter diagram appear close to the line, we note that the correlation coefficient tends to be large. Therefore, the correlation coefficient and the standard error of the estimate are inversely related. As the strength of a linear relationship between two variables increases, the correlation coefficient increases and the standard error of the estimate decreases.

We also noted that the square of the correlation coefficient is the coefficient of determination. The coefficient of determination measures the percentage of the variation in  $Y$  that is explained by the variation in  $X$ .

A convenient vehicle for showing the relationship among these three measures is an ANOVA table. See the highlighted portion of the spreadsheet below. This table is similar to the analysis of variance table developed in Chapter 12. In that chapter, the total variation was divided into two components: variation due to the *treatments* and variation due to *random error*. The concept is similar in regression analysis. The total variation is divided into two components: (1) variation explained by the *regression* (explained by the independent variable) and (2) the *error, or residual*. This is the unexplained variation. These three sources of variance (total, regression, and residual) are identified in the first column of the spreadsheet ANOVA table. The column headed “*df*” refers to the degrees of freedom associated with each category. The total number of degrees of freedom is  $n - 1$ . The number of degrees of freedom in the regression is 1 because there is only one independent variable. The number of degrees of freedom associated with the error term is  $n - 2$ . The term *SS* located in the middle of the ANOVA table refers to the sum of squares. You should note that the total degrees of freedom are equal to the sum of the regression and residual (error) degrees of freedom, and the total sum of squares is equal to the sum of the regression and residual (error) sum of squares. This is true for any ANOVA table.

|    | A                    | B               | C                | D | E                     | F            | G              | H        | I       | J              |
|----|----------------------|-----------------|------------------|---|-----------------------|--------------|----------------|----------|---------|----------------|
| 1  | Sales Representative | Sales Calls (x) | Copiers Sold (y) |   | SUMMARY OUTPUT        |              |                |          |         |                |
| 2  | Brian Virost         | 96              | 41               |   | Regression Statistics |              |                |          |         |                |
| 3  | Carlos Ramirez       | 40              | 41               |   | Multiple R            | 0.865        |                |          |         |                |
| 4  | Carol Saia           | 104             | 51               |   | R Square              | 0.748        |                |          |         |                |
| 5  | Greg Fish            | 128             | 60               |   | Adjusted R Square     | 0.728        |                |          |         |                |
| 6  | Jeff Hall            | 164             | 61               |   | Standard Error        | 6.720        |                |          |         |                |
| 7  | Mark Reynolds        | 76              | 29               |   | Observations          | 15           |                |          |         |                |
| 8  | Meryl Rumsey         | 72              | 39               |   |                       |              |                |          |         |                |
| 9  | Mike Kiel            | 80              | 50               |   | ANOVA                 |              |                |          |         |                |
| 10 | Ray Snarsky          | 36              | 28               |   |                       | df           | SS             | MS       | F       | Significance F |
| 11 | Rich Niles           | 84              | 43               |   | Regression            | 1            | 1738.890       | 1738.890 | 38.503  | 0.000          |
| 12 | Ron Broderick        | 180             | 70               |   | Residual              | 13           | 587.110        | 45.162   |         |                |
| 13 | Sal Spina            | 132             | 56               |   | Total                 | 14           | 2326           |          |         |                |
| 14 | Sani Jones           | 120             | 45               |   |                       |              |                |          |         |                |
| 15 | Susan Welch          | 44              | 31               |   |                       | Coefficients | Standard Error | t Stat   | P-value |                |
| 16 | Tom Keller           | 84              | 30               |   | Intercept             | 19.980       | 4.390          | 4.552    | 0.001   |                |
| 17 |                      |                 |                  |   | Sales Calls (x)       | 0.261        | 0.042          | 6.205    | 0.000   |                |

Source: Microsoft Excel

The ANOVA sums of squares are:

$$\text{Regression Sum of Squares} = \text{SSR} = \sum(\hat{y} - \bar{y})^2 = 1738.89$$

$$\text{Residual or Error Sum of Squares} = \text{SSE} = \sum(y - \hat{y})^2 = 587.11$$

$$\text{Total Sum of Squares} = \text{SS Total} = \sum(y - \bar{y})^2 = 2326.0$$

Recall that the coefficient of determination is defined as the percentage of the total variation (SS Total) explained by the regression equation (SSR). Using the ANOVA table, the reported value of *R*-square can be validated.

**COEFFICIENT OF DETERMINATION**

$$r^2 = \frac{\text{SSR}}{\text{SS Total}} = 1 - \frac{\text{SSE}}{\text{SS Total}} \tag{13-8}$$

Using the values from the ANOVA table, the coefficient of determination is  $1738.89/2326.0 = 0.748$ . Therefore, the more variation of the dependent variable (SS Total) explained by the independent variable (SSR), the higher the coefficient of determination.

We can also express the coefficient of determination in terms of the error or residual variation:

$$r^2 = 1 - \frac{SSE}{SS \text{ Total}} = 1 - \frac{587.11}{2326.0} = 1 - 0.252 = 0.748$$

As illustrated in formula (13–8), the coefficient of determination and the residual or error sum of squares are inversely related. The higher the unexplained or error variation as a percentage of the total variation, the lower is the coefficient of determination. In this case, 25.2% of the total variation in the dependent variable is error or residual variation.

The final observation that relates the correlation coefficient, the coefficient of determination, and the standard error of estimate is to show the relationship between the standard error of estimate and SSE. By substituting [SSE Residual or Error Sum of Squares = SSE =  $\Sigma(y - \hat{y})^2$ ] into the formula for the standard error of estimate, we find:

**STANDARD ERROR  
OF ESTIMATE**

$$s_{y \cdot x} = \sqrt{\frac{SSE}{n - 2}}$$

**(13–9)**

Note that  $s_{y \cdot x}$  can also be computed using the residual mean square from the ANOVA table.

**STANDARD ERROR  
OF THE ESTIMATE**

$$s_{y \cdot x} = \sqrt{\text{Residual mean square}}$$

**(13–10)**

To summarize, regression analysis provides two statistics to evaluate the predictive ability of a regression equation: the standard error of the estimate and the coefficient of determination. When reporting the results of a regression analysis, the findings must be clearly explained, especially when using the results to make predictions of the dependent variable. The report must always include a statement regarding the coefficient of determination so that the relative precision of the prediction is known to the reader of the report. Objective reporting of statistical analysis is required so that the readers can make their own decisions.

## EXERCISES

29. Given the following ANOVA table:

| Source     | df | SS     | MS     | F     |
|------------|----|--------|--------|-------|
| Regression | 1  | 1000.0 | 1000.0 | 26.00 |
| Error      | 13 | 500.0  | 38.46  |       |
| Total      | 14 | 1500.0 |        |       |

- Determine the coefficient of determination.
  - Assuming a direct relationship between the variables, what is the correlation coefficient?
  - Determine the standard error of estimate.
30. On the first statistics exam, the coefficient of determination between the hours studied and the grade earned was 80%. The standard error of estimate was 10. There were 20 students in the class. Develop an ANOVA table for the regression analysis of hours studied as a predictor of the grade earned on the first statistics exam.

**LO13-6**

Calculate and interpret confidence and prediction intervals.

**STATISTICS IN ACTION**

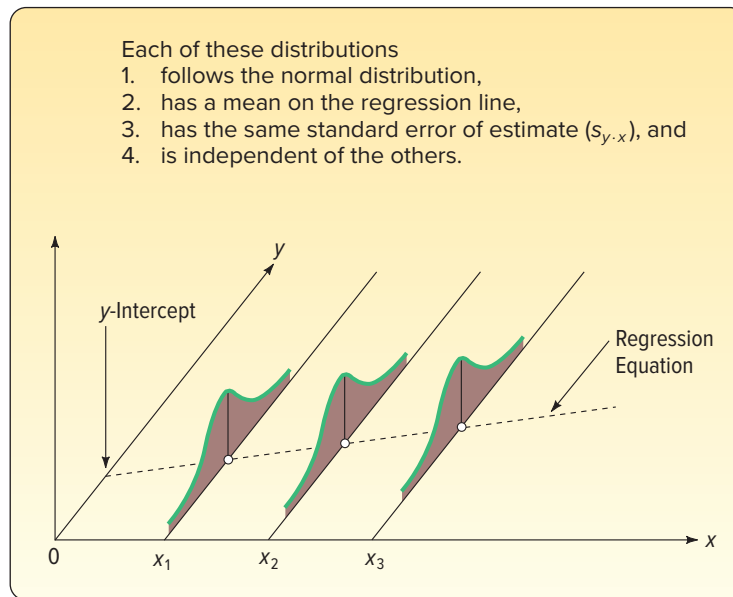
Studies indicate that for both men and women, those who are considered good looking earn higher wages than those who are not. In addition, for men there is a correlation between height and salary. For each additional inch of height, a man can expect to earn an additional \$250 per year. So a man 6'6" tall receives a \$3,000 "stature" bonus over his 5'6" counterpart. Being overweight or underweight is also related to earnings, particularly among women. A study of young women showed the heaviest 10% earned about 6% less than their lighter counterparts.

## INTERVAL ESTIMATES OF PREDICTION

The standard error of estimate and the coefficient of determination are two statistics that provide an overall evaluation of the ability of a regression equation to predict a dependent variable. Another way to report the ability of a regression equation to predict is specific to a stated value of the independent variable. For example, we can predict the number of copiers sold ( $y$ ) for a selected value of number of sales calls made ( $x$ ). In fact, we can calculate a confidence interval for the predicted value of the dependent variable for a selected value of the independent variable.

### Assumptions Underlying Linear Regression

Before we present the confidence intervals, the assumptions for properly applying linear regression should be reviewed. Chart 13–13 illustrates these assumptions.



**CHART 13–13** Regression Assumptions Shown Graphically

1. For each value of  $x$ , there are corresponding  $y$  values. These  $y$  values follow the normal distribution.
2. The means of these normal distributions lie on the regression line.
3. The standard deviations of these normal distributions are all the same. The best estimate we have of this common standard deviation is the standard error of estimate ( $s_{y \cdot x}$ ).
4. The  $y$  values are statistically independent. This means that in selecting a sample, a particular  $x$  does not depend on any other value of  $x$ . This assumption is particularly important when data are collected over a period of time. In such situations, the errors for a particular time period are often correlated with those of other time periods.

Recall from Chapter 7 that if the values follow a normal distribution, then the mean plus or minus one standard deviation will encompass 68% of the observations, the mean plus or minus two standard deviations will encompass 95% of the observations, and

the mean plus or minus three standard deviations will encompass virtually all of the observations. The same relationship exists between the predicted values  $\hat{y}$  and the standard error of estimate ( $s_{y \cdot x}$ ).

1.  $\hat{y} \pm s_{y \cdot x}$  will include the middle 68% of the observations.
2.  $\hat{y} \pm 2s_{y \cdot x}$  will include the middle 95% of the observations.
3.  $\hat{y} \pm 3s_{y \cdot x}$  will include virtually all the observations.

We can now relate these assumptions to North American Copier Sales, where we studied the relationship between the number of sales calls and the number of copiers sold. If we drew a parallel line 6.72 units above the regression line and another 6.72 units below the regression line, about 68% of the points would fall between the two lines. Similarly, a line 13.44 [ $2s_{y \cdot x} = 2(6.72)$ ] units above the regression line and another 13.44 units below the regression line should include about 95% of the data values.

As a rough check, refer to column E in the Excel spreadsheet appearing on page 385. Four of the 15 deviations exceed one standard error of estimate. That is, the deviations of Carlos Ramirez, Mark Reynolds, Mike Kiel, and Tom Keller all exceed 6.72 (one standard error). All values are less than 13.44 units away from the regression line. In short, 11 of the 15 deviations are within one standard error and all are within two standard errors. That is a fairly good result for a relatively small sample.

## Constructing Confidence and Prediction Intervals

When using a regression equation, two different predictions can be made for a selected value of the independent variable. The differences are subtle but very important and are related to the assumptions stated in the last section. Recall that for any selected value of the independent variable ( $X$ ), the dependent variable ( $Y$ ) is a random variable that is normally distributed with a mean  $\hat{Y}$ . Each distribution of  $Y$  has a standard deviation equal to the regression analysis's standard error of estimate.

The first interval estimate is called a **confidence interval**. This is used when the regression equation is used to predict the mean value of  $Y$  for a given value of  $x$ . For example, we would use a confidence interval to estimate the mean salary of all executives in the retail industry based on their years of experience. To determine the confidence interval for the mean value of  $y$  for a given  $x$ , the formula is:

**CONFIDENCE INTERVAL FOR  
THE MEAN OF Y, GIVEN X**

$$\hat{y} \pm ts_{y \cdot x} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}} \quad (13-11)$$

The second interval estimate is called a prediction interval. This is used when the regression equation is used to predict an individual  $y$  for a given value of  $x$ . For example, we would estimate the salary of a particular retail executive who has 20 years of experience. To calculate a prediction interval, formula (13-11) is modified by adding a 1 under the radical. To determine the prediction interval for an estimate of an individual for a given  $x$ , the formula is:

**PREDICTION INTERVAL  
FOR Y, GIVEN X**

$$\hat{y} \pm ts_{y \cdot x} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}} \quad (13-12)$$

### EXAMPLE

We return to the North American Copier Sales illustration. Determine a 95% confidence interval for all sales representatives who make 50 calls, and determine a prediction interval for Sheila Baker, a West Coast sales representative who made 50 calls.

### SOLUTION

We use formula (13–11) to determine a confidence level. Table 13–4 includes the necessary totals and a repeat of the information in Table 13–2.

**TABLE 13–4** Determining Confidence and Prediction Intervals

| Sales Representative | Sales Calls ( $x$ ) | Copiers Sold ( $y$ ) | $(x - \bar{x})$ | $(x - \bar{x})^2$ |
|----------------------|---------------------|----------------------|-----------------|-------------------|
| Brian Virost         | 96                  | 41                   | 0               | 0                 |
| Carlos Ramirez       | 40                  | 41                   | -56             | 3,136             |
| Carol Saia           | 104                 | 51                   | 8               | 64                |
| Greg Fish            | 128                 | 60                   | 32              | 1,024             |
| Jeff Hall            | 164                 | 61                   | 68              | 4,624             |
| Mark Reynolds        | 76                  | 29                   | -20             | 400               |
| Meryl Rumsey         | 72                  | 39                   | -24             | 576               |
| Mike Kiel            | 80                  | 50                   | -16             | 256               |
| Ray Snarsky          | 36                  | 28                   | -60             | 3,600             |
| Rich Niles           | 84                  | 43                   | -12             | 144               |
| Ron Broderick        | 180                 | 70                   | 84              | 7,056             |
| Sal Spina            | 132                 | 56                   | 36              | 1,296             |
| Sani Jones           | 120                 | 45                   | 24              | 576               |
| Susan Welch          | 44                  | 31                   | -52             | 2,704             |
| Tom Keller           | 84                  | 30                   | -12             | 144               |
| Total                | 1,440               | 675                  | 0               | 25,600            |

The first step is to determine the number of copiers we expect a sales representative to sell if he or she makes 50 calls. It is 33.0032, found by

$$\hat{y} = 19.9632 + 0.2608x = 19.9632 + 0.2608(50) = 33.0032$$

To find the  $t$  value, we need to first know the number of degrees of freedom. In this case, the degrees of freedom are  $n - 2 = 15 - 2 = 13$ . We set the confidence level at 95%. To find the value of  $t$ , move down the left-hand column of Appendix B.5 to 13 degrees of freedom, then move across to the column with the 95% level of confidence. The value of  $t$  is 2.160.

In the previous section, we calculated the standard error of estimate to be 6.720. We let  $x = 50$ , and from Table 13–4, the mean number of sales calls is 96.0 ( $1,440/15$ ) and  $\Sigma(x - \bar{x})^2 = 25,600$ . Inserting these values in formula (13–11), we can determine the confidence interval.

$$\begin{aligned} \text{Confidence Interval} &= \hat{y} \pm t_{s_{y \cdot x}} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma(x - \bar{x})^2}} \\ &= 33.0032 \pm 2.160(6.720) \sqrt{\frac{1}{15} + \frac{(50 - 96)^2}{25,600}} \\ &= 33.0032 \pm 5.6090 \end{aligned}$$

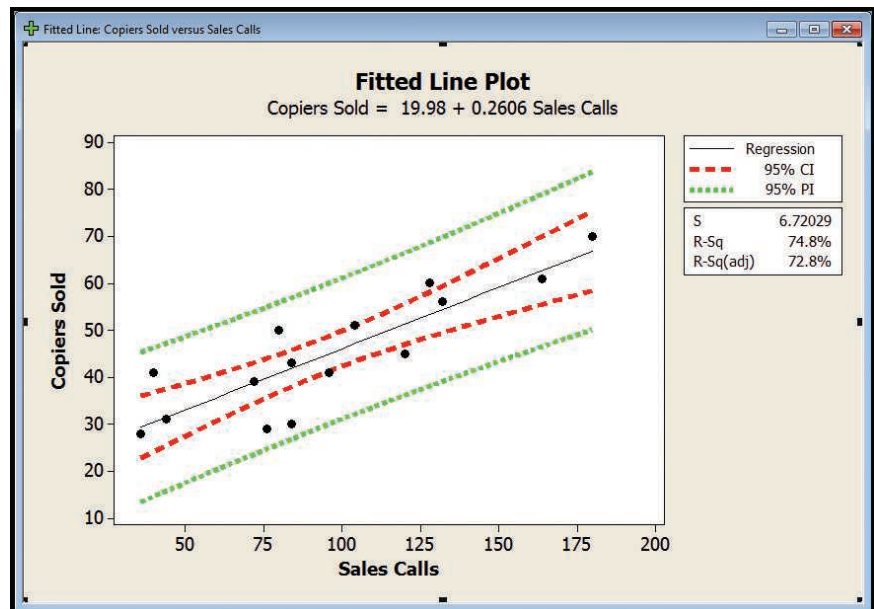
Thus, the 95% confidence interval for all sales representatives who make 50 calls is from 27.3942 up to 38.6122. To interpret, let's round the values. If a sales representative makes 50 calls, he or she can expect to sell 33 copiers. It is likely the sales will range from 27.4 to 38.6 copiers.

Suppose we want to estimate the number of copiers sold by Sheila Baker, who made 50 sales calls. Using formula (13–12), the 95% prediction interval is determined as follows:

$$\begin{aligned} \text{Prediction Interval} &= \hat{y} \pm t_{s_{y \cdot x}} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}} \\ &= 33.0032 \pm 2.160(6.720) \sqrt{1 + \frac{1}{15} + \frac{(50 - 96)^2}{25,600}} \\ &= 33.0032 \pm 15.5612 \end{aligned}$$

Thus, the interval is from 17.442 up to 48.5644 copiers. We conclude that the number of office machines sold will be between about 17.4 and 48.6 for a particular sales representative, such as Sheila Baker, who makes 50 calls. This interval is quite large. It is much larger than the confidence interval for all sales representatives who made 50 calls. It is logical, however, that there should be more variation in the sales estimate for an individual than for a group.

The following Minitab graph shows the relationship between the least squares regression line (in the center), the confidence interval (shown in crimson), and the prediction interval (shown in green). The bands for the prediction interval are always further from the regression line than those for the confidence interval. Also, as the values of  $x$  move away from the mean number of calls (96) in either direction, the confidence interval and prediction interval bands widen. This is caused by the numerator of the right-hand term under the radical in formulas (13–11) and (13–12). That is, as the term increases, the widths of the confidence interval and the prediction interval also increase. To put it another way, there is less precision in our estimates as we move away, in either direction, from the mean of the independent variable.



Source: Minitab



We wish to emphasize again the distinction between a confidence interval and a prediction interval. A confidence interval refers to the mean of all cases for a given value of  $x$  and is computed by formula (13–11). A prediction interval refers to a particular, single case for a given value of  $x$  and is computed using formula (13–12). The prediction interval will always be wider because of the extra 1 under the radical in the second equation.

## SELF-REVIEW 13–6



Refer to the sample data in Self-Review 13–1, where the owner of Haverty's Furniture was studying the relationship between sales and the amount spent on advertising. The advertising expense and sales revenue, both in millions of dollars, for the last 4 months are repeated below.

| Month     | Advertising Expense<br>(\$ million) | Sales Revenue<br>(\$ million) |
|-----------|-------------------------------------|-------------------------------|
| July      | 2                                   | 7                             |
| August    | 1                                   | 3                             |
| September | 3                                   | 8                             |
| October   | 4                                   | 10                            |

The regression equation was computed to be  $\hat{y} = 1.5 + 2.2x$ , and the standard error was 0.9487. Both variables are reported in millions of dollars. Determine the 90% confidence interval for the typical month in which \$3 million was spent on advertising.

## EXERCISES

31. Refer to Exercise 13.
  - a. Determine the .95 confidence interval for the mean predicted when  $x = 7$ .
  - b. Determine the .95 prediction interval for an individual predicted when  $x = 7$ .
32. Refer to Exercise 14.
  - a. Determine the .95 confidence interval for the mean predicted when  $x = 7$ .
  - b. Determine the .95 prediction interval for an individual predicted when  $x = 7$ .
33. Refer to Exercise 15.
  - a. Determine the .95 confidence interval, in thousands of kilowatt-hours, for the mean of all six-room homes.
  - b. Determine the .95 prediction interval, in thousands of kilowatt-hours, for a particular six-room home.
34. Refer to Exercise 16.
  - a. Determine the .95 confidence interval, in thousands of dollars, for the mean of all sales personnel who make 40 contacts.
  - b. Determine the .95 prediction interval, in thousands of dollars, for a particular salesperson who makes 40 contacts.

## TRANSFORMING DATA

### LO13-7

Use a log function to transform a nonlinear relationship.

Regression analysis describes the relationship between two variables. A requirement is that this relationship be linear. The same is true of the correlation coefficient. It measures the strength of a linear relationship between two variables. But what if the relationship is not linear? The remedy is to rescale one or both of the variables so the new relationship is linear. For example, instead of using the actual values of the dependent variable,  $y$ , we would create a new dependent variable by computing the log to the

base 10 of  $y$ ,  $\text{Log}(y)$ . This calculation is called a transformation. Other common transformations include taking the square root, taking the reciprocal, or squaring one or both of the variables.

Thus, two variables could be closely related, but their relationship is not linear. Be cautious when you are interpreting the correlation coefficient or a regression equation. These statistics may indicate there is no linear relationship, but there could be a relationship of some other nonlinear or curvilinear form. The following example explains the details.

### EXAMPLE

GroceryLand Supermarkets is a regional grocery chain with over 300 stores located in the midwestern United States. The corporate director of marketing for GroceryLand wishes to study the effect of price on the weekly sales of two-liter bottles of their private-brand diet cola. The objectives of the study are:

1. To determine whether there is a relationship between selling price and weekly sales. Is this relationship direct or indirect? Is it strong or weak?
2. To determine the effect of price increases or decreases on sales. Can we effectively forecast sales based on the price?

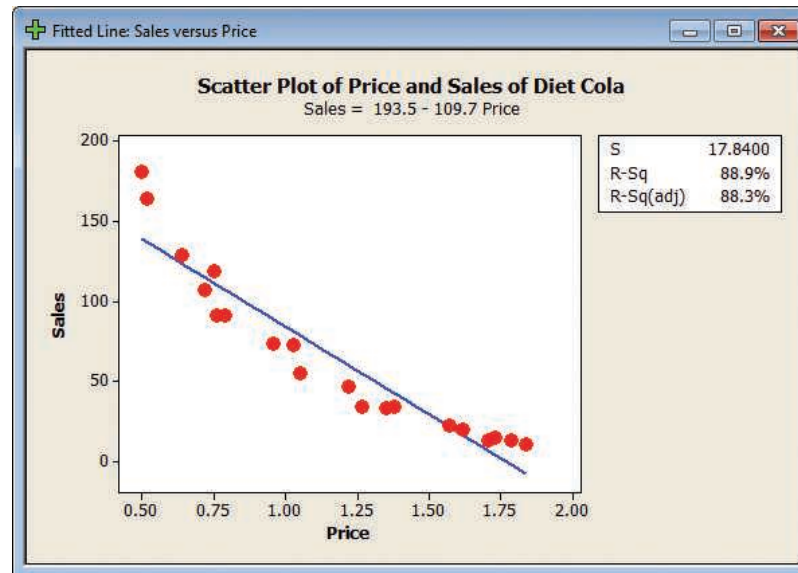
### SOLUTION

To begin the project, the marketing director meets with the vice president of sales and other company staff members. They decide that it would be reasonable to price the two-liter bottle of their private-brand diet cola from \$0.50 up to \$2.00. To collect the data needed to analyze the relationship between price and sales, the marketing director selects a random sample of 20 stores and then randomly assigns a selling price for the two-liter bottle of diet cola between \$0.50 and \$2.00 to each selected store. The director contacts each of the 20 store managers included in the study to tell them the selling price and ask them to report the sales for the product at the end of the week. The results are reported below. For example, store number A-17 sold 181 two-liter bottles of diet cola at \$0.50 each.

| GroceryLand Sales and Price Data |       |       | GroceryLand Sales and Price Data |       |       |
|----------------------------------|-------|-------|----------------------------------|-------|-------|
| Store Number                     | Price | Sales | Store Number                     | Price | Sales |
| A-17                             | 0.50  | 181   | A-30                             | 0.76  | 91    |
| A-121                            | 1.35  | 33    | A-127                            | 1.79  | 13    |
| A-227                            | 0.79  | 91    | A-266                            | 1.57  | 22    |
| A-135                            | 1.71  | 13    | A-117                            | 1.27  | 34    |
| A-6                              | 1.38  | 34    | A-132                            | 0.96  | 74    |
| A-282                            | 1.22  | 47    | A-120                            | 0.52  | 164   |
| A-172                            | 1.03  | 73    | A-272                            | 0.64  | 129   |
| A-296                            | 1.84  | 11    | A-120                            | 1.05  | 55    |
| A-143                            | 1.73  | 15    | A-194                            | 0.72  | 107   |
| A-66                             | 1.62  | 20    | A-105                            | 0.75  | 119   |

To examine the relationship between Price and Sales, we use regression analysis setting *Price* as the independent variable and *Sales* as the dependent

variable. The analysis will provide important information about the relationship between the variables. The analysis is summarized in the following Minitab output.



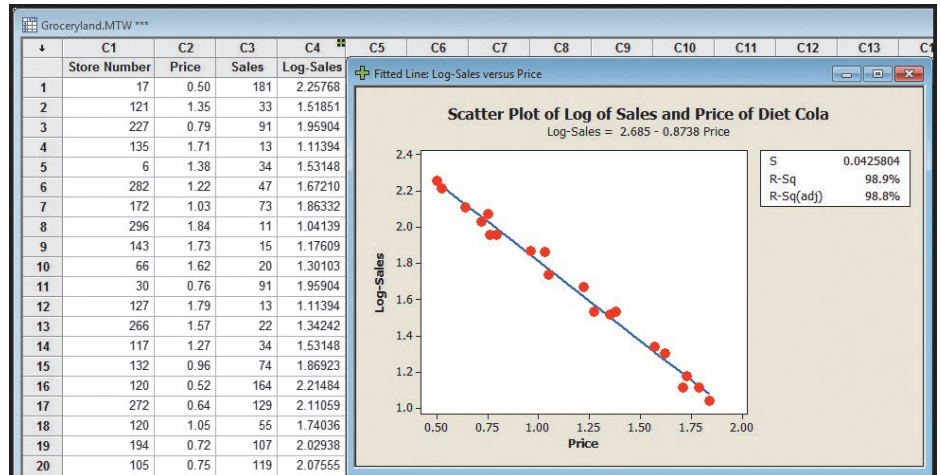
Source: Minitab

From the output, we can make these conclusions:

1. The relationship between the two variables is inverse or indirect. As the *Price* of the cola increases, the *Sales* of the product decrease. Given basic economic theory of price and demand, this is expected.
2. There is a strong relationship between the two variables. The coefficient of determination is 88.9%. So 88.9% of the variation in *Sales* is accounted for by the variation in *Price*. From the coefficient of determination, we can compute the correlation coefficient as the square root of the coefficient of determination. The correlation coefficient is the square root of 0.889, or 0.943. The sign of the correlation coefficient is negative because sales are inversely related to price. Therefore, the correlation coefficient is  $-0.943$ .
3. Before continuing our summary of conclusions, we should look carefully at the scatter diagram and the plot of the regression line. The assumption of a linear relationship is tenuous. If the relationship is linear, the data points should be distributed both above and below the line over the entire range of the independent variable. However, for the highest and lowest prices, the data points are above the regression line. For the selling prices in the middle, most of the data points are below the regression line. So the linear regression equation does not effectively describe the relationship between *Price* and *Sales*. A transformation of the data is needed to create a linear relationship.

By transforming one of the variables, we may be able to change the nonlinear relationship between the variables to a linear relationship. Of the possible choices, the director of marketing decides to transform the dependent variable, *Sales*, by taking the logarithm to the base 10 of each *Sales* value. Note the new variable, *Log-Sales*, in the following analysis. Now, the regression analysis uses *Log-Sales* as

the dependent variable and *Price* as the independent variable. This analysis is reported below.



Source: Minitab

What can we conclude from the regression analysis using the transformation of the dependent variable *Sales*?

1. By transforming the dependent variable, *Sales*, we increase the coefficient of determination from 0.889 to 0.989. So *Price* now explains nearly all of the variation in *Log-Sales*.
2. Compare this result with the scatter diagram before we transformed the dependent variable. The transformed data seem to fit the linear relationship requirement much better. Observe that the data points are both above and below the regression line over the range of *Price*.
3. The regression equation is  $\hat{y} = 2.685 - 0.8738x$ . The sign of the slope value is negative, confirming the inverse association between the variables. We can use the new equation to estimate sales and study the effect of changes in price. For example, if we decided to sell the two-liter bottle of diet cola for \$1.25, the predicted *Log-Sales* is:

$$\hat{y} = 2.685 - 0.8738x = 2.685 - 0.8738(1.25) = 1.593$$

Remember that the regression equation now predicts the log, base 10, of *Sales*. Therefore, we must undo the transformation by taking the antilog of 1.593, which is  $10^{1.593}$ , or 39.174. So, if we price the two-liter diet cola product at \$1.25, the predicted weekly sales are 39 bottles. If we increase the price to \$2.00, the regression equation would predict a value of .9374. Taking the antilog,  $10^{.9374}$ , the predicted sales decrease to 8.658, or, rounding, 9 two-liter bottles per week. Clearly, as price increases, sales decrease. This relationship will be very helpful to GroceryLand when making pricing decisions for this product.

## EXERCISES

35. **FILE** Given the following sample of five observations, develop a scatter diagram, using  $x$  as the independent variable and  $y$  as the dependent variable, and compute the correlation coefficient. Does the relationship between the variables appear to be linear? Try squaring the  $x$  variable and then develop a scatter diagram and determine the correlation coefficient. Summarize your analysis.

|     |    |     |     |   |     |
|-----|----|-----|-----|---|-----|
| $x$ | −8 | −16 | 12  | 2 | 18  |
| $y$ | 58 | 247 | 153 | 3 | 341 |

36. **FILE** Every April, The Masters—one of the most prestigious golf tournaments on the PGA golf tour—is played in Augusta, Georgia. In 2016, 55 players received prize money. The 2016 winner, Danny Willett, earned a prize of \$1,800,000. Jordan Spieth and Lee Westwood tied for second place, earning \$880,000. Two amateur players finished “in the money,” but they could not accept the prize money. They are not included in the data. The data are briefly summarized below. Each player has three corresponding variables: finishing position, score, and prize (in dollars). We want to study the relationship between score and prize.

| Position | Player           | Score | Prize       |
|----------|------------------|-------|-------------|
| 1        | Danny Willett    | 283   | \$1,800,000 |
| 2(tied)  | Jordan Spieth    | 286   | \$880,000   |
| 2(tied)  | Lee Westwood     | 286   | \$880,000   |
| 4(tied)  | Paul Casey       | 287   | \$413,333   |
| 4(tied)  | J.B. Holmes      | 287   | \$413,333   |
| 4(tied)  | Dustin Johnson   | 287   | \$413,333   |
| ..       | ..               | ..    | ..          |
| ..       | ..               | ..    | ..          |
| ..       | ..               | ..    | ..          |
| 52(tied) | Keegan Bradley   | 301   | \$24,900    |
| 52(tied) | Larry Mize       | 301   | \$24,900    |
| 54       | Hunter Mahan     | 302   | \$24,000    |
| 55(tied) | Kevin Na         | 303   | \$23,400    |
| 55(tied) | Cameron Smith    | 303   | \$23,400    |
| 57       | Thongchai Jaidee | 307   | \$23,000    |

- Using *Score* as the independent variable and *Prize* as the dependent variable, develop a scatter diagram. Does the relationship appear to be linear? Does it seem reasonable that as *Score*, increases the *Prize* decreases?
- What percentage of the variation in the dependent variable, *Prize*, is accounted for by the independent variable, *Score*?
- Calculate a new variable, *Log-Prize*, computing the log to the base 10 of *Prize*. Draw a scatter diagram with *Log-Prize* as the dependent variable and *Score* as the independent variable.
- Develop a regression equation and compute the coefficient of determination using *Log-Prize* as the dependent variable.
- Compare the coefficient of determination in parts (b) and (d). What do you conclude?
- Write out the regression equation developed in part (d). If a player shot an even par score of 288 for the four rounds, how much would you expect that player to earn?

## CHAPTER SUMMARY

- A scatter diagram is a graphic tool used to portray the relationship between two variables.
  - The dependent variable is scaled on the Y-axis and is the variable being estimated.
  - The independent variable is scaled on the X-axis and is the variable used as the predictor.
- The correlation coefficient measures the strength of the linear association between two variables.
  - Both variables must be at least the interval scale of measurement.
  - The correlation coefficient can range from  $-1.00$  to  $1.00$ .

- C. If the correlation between the two variables is 0, there is no association between them.
- D. A value of 1.00 indicates perfect positive correlation, and a value of  $-1.00$  indicates perfect negative correlation.
- E. A positive sign means there is a direct relationship between the variables, and a negative sign means there is an inverse relationship.
- F. It is designated by the letter  $r$  and found by the following equation:

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y} \quad (13-1)$$

- G. To test a hypothesis that a population correlation is different from 0, we use the following statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{with } n - 2 \text{ degrees of freedom} \quad (13-2)$$

III. In regression analysis, we estimate one variable based on another variable.

- A. The variable being estimated is the dependent variable.
- B. The variable used to make the estimate or predict the value is the independent variable.
  1. The relationship between the variables is linear.
  2. Both the independent and the dependent variables must be interval or ratio scale.
  3. The least squares criterion is used to determine the regression equation.

IV. The least squares regression line is of the form  $\hat{y} = a + bx$ .

- A.  $\hat{y}$  is the estimated value of  $y$  for a selected value of  $x$ .
- B.  $a$  is the constant or intercept.
  1. It is the value of  $\hat{y}$  when  $x = 0$ .
  2.  $a$  is computed using the following equation.

$$a = \bar{y} - b\bar{x} \quad (13-5)$$

- C.  $b$  is the slope of the fitted line.
  1. It shows the amount of change in  $\hat{y}$  for a change of one unit in  $x$ .
  2. A positive value for  $b$  indicates a direct relationship between the two variables. A negative value indicates an inverse relationship.
  3. The sign of  $b$  and the sign of  $r$ , the correlation coefficient, are always the same.
  4.  $b$  is computed using the following equation.

$$b = r \left( \frac{s_y}{s_x} \right) \quad (13-4)$$

- D.  $x$  is the value of the independent variable.

V. For a regression equation, the slope is tested for significance.

- A. We test the hypothesis that the slope of the line in the population is 0.
  1. If we do not reject the null hypothesis, we conclude there is no relationship between the two variables.
  2. The test is equivalent to the test for the correlation coefficient.
- B. When testing the null hypothesis about the slope, the test statistic is with  $n - 2$  degrees of freedom:

$$t = \frac{b - 0}{s_b} \quad (13-6)$$

VI. The standard error of estimate measures the variation around the regression line.

- A. It is in the same units as the dependent variable.
- B. It is based on squared deviations from the regression line.
- C. Small values indicate that the points cluster closely about the regression line.
- D. It is computed using the following formula.

$$s_{y \cdot x} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}} \quad (13-7)$$

- VII.** The coefficient of determination is the proportion of the variation of a dependent variable explained by the independent variable.
- It ranges from 0 to 1.0.
  - It is the square of the correlation coefficient.
  - It is found from the following formula.

$$r^2 = \frac{SSR}{SS \text{ Total}} = 1 - \frac{SSE}{SS \text{ Total}} \quad (13-8)$$

- VIII.** Inference about linear regression is based on the following assumptions.
- For a given value of  $x$ , the values of  $y$  are normally distributed about the line of regression.
  - The standard deviation of each of the normal distributions is the same for all values of  $x$  and is estimated by the standard error of estimate.
  - The deviations from the regression line are independent, with no pattern to the size or direction.
- IX.** There are two types of interval estimates.
- In a confidence interval, the mean value of  $y$  is estimated for a given value of  $x$ .
    - It is computed from the following formula.

$$\hat{y} \pm t_{s_{y \cdot x}} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}} \quad (13-11)$$

- The width of the interval is affected by the level of confidence, the size of the standard error of estimate, and the size of the sample, as well as the value of the independent variable.
- In a prediction interval, the individual value of  $y$  is estimated for a given value of  $x$ .
    - It is computed from the following formula.

$$\hat{y} \pm t_{s_{y \cdot x}} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}} \quad (13-12)$$

- The difference between formulas (13-11) and (13-12) is the 1 under the radical.
  - The prediction interval will be wider than the confidence interval.
  - The prediction interval is also based on the level of confidence, the size of the standard error of estimate, the size of the sample, and the value of the independent variable.

## PRONUNCIATION KEY

| SYMBOL          | MEANING                                   | PRONUNCIATION        |
|-----------------|---|----------------------|
| $\sum xy$       | Sum of the products of $x$ and $y$        | <i>Sum x y</i>       |
| $\rho$          | Correlation coefficient in the population | <i>Rho</i>           |
| $\hat{y}$       | Estimated value of $Y$                    | <i>y hat</i>         |
| $s_{y \cdot x}$ | Standard error of estimate                | <i>s sub y dot x</i> |
| $r^2$           | Coefficient of determination              | <i>r square</i>      |

## CHAPTER EXERCISES

- A regional commuter airline selected a random sample of 25 flights and found that the correlation between the number of passengers and the total weight, in pounds, of luggage stored in the luggage compartment is 0.94. Using the .05 significance level, can we conclude that there is a positive association between the two variables?
- A sociologist claims that the success of students in college (measured by their GPA) is related to their family's income. For a sample of 20 students, the correlation coefficient is 0.40. Using the .01 significance level, can we conclude that there is a positive correlation between the variables?

39. An Environmental Protection Agency study of 12 automobiles revealed a correlation of 0.47 between engine size and emissions. At the .01 significance level, can we conclude that there is a positive association between these variables? What is the  $p$ -value? Interpret.
40. **FILE** A suburban hotel derives its revenue from its hotel and restaurant operations. The owners are interested in the relationship between the number of rooms occupied on a nightly basis and the revenue per day in the restaurant. Below is a sample of 25 days (Monday through Thursday) from last year showing the restaurant income and number of rooms occupied.

| Day | Revenue | Occupied | Day | Revenue | Occupied |
|-----|---------|----------|-----|---------|----------|
| 1   | \$1,452 | 23       | 14  | \$1,425 | 27       |
| 2   | 1,361   | 47       | 15  | 1,445   | 34       |
| 3   | 1,426   | 21       | 16  | 1,439   | 15       |
| 4   | 1,470   | 39       | 17  | 1,348   | 19       |
| 5   | 1,456   | 37       | 18  | 1,450   | 38       |
| 6   | 1,430   | 29       | 19  | 1,431   | 44       |
| 7   | 1,354   | 23       | 20  | 1,446   | 47       |
| 8   | 1,442   | 44       | 21  | 1,485   | 43       |
| 9   | 1,394   | 45       | 22  | 1,405   | 38       |
| 10  | 1,459   | 16       | 23  | 1,461   | 51       |
| 11  | 1,399   | 30       | 24  | 1,490   | 61       |
| 12  | 1,458   | 42       | 25  | 1,426   | 39       |
| 13  | 1,537   | 54       |     |         |          |

Use a statistical software package to answer the following questions.

- Does the revenue seem to increase as the number of occupied rooms increases? Draw a scatter diagram to support your conclusion.
  - Determine the correlation coefficient between the two variables. Interpret the value.
  - Is it reasonable to conclude that there is a positive relationship between revenue and occupied rooms? Use the .10 significance level.
  - What percent of the variation in revenue in the restaurant is accounted for by the number of rooms occupied?
41. **FILE** The table below shows the number of cars (in millions) sold in the United States for various years and the percent of those cars manufactured by GM.

| Year | Cars Sold (millions) | Percent GM | Year | Cars Sold (millions) | Percent GM |
|------|----------------------|------------|------|----------------------|------------|
| 1950 | 6.0                  | 50.2       | 1985 | 15.4                 | 40.1       |
| 1955 | 7.8                  | 50.4       | 1990 | 13.5                 | 36.0       |
| 1960 | 7.3                  | 44.0       | 1995 | 15.5                 | 31.7       |
| 1965 | 10.3                 | 49.9       | 2000 | 17.4                 | 28.6       |
| 1970 | 10.1                 | 39.5       | 2005 | 16.9                 | 26.9       |
| 1975 | 10.8                 | 43.1       | 2010 | 11.6                 | 19.1       |
| 1980 | 11.5                 | 44.0       | 2015 | 17.5                 | 17.6       |

Use a statistical software package to answer the following questions.

- Is the number of cars sold directly or indirectly related to GM's percentage of the market? Draw a scatter diagram to show your conclusion.
  - Determine the correlation coefficient between the two variables. Interpret the value.
  - Is it reasonable to conclude that there is a negative association between the two variables? Use the .01 significance level.
  - How much of the variation in GM's market share is accounted for by the variation in cars sold?
42. For a sample of 40 large U.S. cities, the correlation between the mean number of square feet per office worker and the mean monthly rental rate in the central business district is  $-0.363$ . At the .05 significance level, can we conclude that there is a negative association between the two variables?



43. **FILE** For each of the 32 National Football League teams, the number of points scored and allowed during the 2016 season are shown below.

| TEAM         | Conference | PTS<br>Scored | PTS<br>Allowed | TEAM          | Conference | PTS<br>Scored | PTS<br>Allowed |
|--------------|------------|---------------|----------------|---------------|------------|---------------|----------------|
| Baltimore    | AFC        | 343           | 321            | Arizona       | NFC        | 418           | 362            |
| Buffalo      | AFC        | 399           | 378            | Atlanta       | NFC        | 540           | 406            |
| Cincinnati   | AFC        | 325           | 315            | Carolina      | NFC        | 369           | 402            |
| Cleveland    | AFC        | 264           | 452            | Chicago       | NFC        | 279           | 399            |
| Denver       | AFC        | 333           | 297            | Dallas        | NFC        | 421           | 306            |
| Houston      | AFC        | 279           | 328            | Detroit       | NFC        | 346           | 358            |
| Indianapolis | AFC        | 411           | 392            | Green Bay     | NFC        | 432           | 388            |
| Jacksonville | AFC        | 318           | 400            | Los Angeles   | NFC        | 224           | 394            |
| Kansas City  | AFC        | 389           | 311            | Minnesota     | NFC        | 327           | 307            |
| Miami        | AFC        | 363           | 380            | NY Giants     | NFC        | 469           | 454            |
| New England  | AFC        | 441           | 250            | New Orleans   | NFC        | 310           | 284            |
| NY Jets      | AFC        | 275           | 409            | Philadelphia  | NFC        | 367           | 331            |
| Oakland      | AFC        | 416           | 385            | San Francisco | NFC        | 309           | 480            |
| Pittsburgh   | AFC        | 399           | 327            | Seattle       | NFC        | 354           | 292            |
| San Diego    | AFC        | 410           | 423            | Tampa Bay     | NFC        | 354           | 369            |
| Tennessee    | AFC        | 381           | 378            | Washington    | NFC        | 396           | 383            |

Assuming these are sample data, answer the following questions. You may use statistical software to assist you.

- What is the correlation coefficient between these variables? Are you surprised the association is negative? Interpret your results.
  - Find the coefficient of determination. What does it say about the relationship?
  - At the .05 significance level, can you conclude there is a negative association between “points scored” and “points allowed”?
  - At the .05 significance level, can you conclude there is a negative association between “points scored” and “points allowed” for each conference?
44. **FILE** The Cotton Mill is an upscale chain of women’s clothing stores, located primarily in the southwest United States. Due to recent success, The Cotton Mill’s top management is planning to expand by locating new stores in other regions of the country. The director of planning has been asked to study the relationship between yearly sales and the store size. As part of the study, the director selects a sample of 25 stores and determines the size of the store in square feet and the sales for last year. The sample data follow. The use of statistical software is suggested.

| Store Size<br>(thousands of<br>square feet) | Sales<br>(millions \$) | Store Size<br>(thousands of<br>square feet) | Sales<br>(millions \$) |
|---|------------------------|---|------------------------|
| 3.7   | 9.18                   | 0.4   | 0.55                   |
| 2.0   | 4.58                   | 4.2   | 7.56                   |
| 5.0   | 8.22                   | 3.1   | 2.23                   |
| 0.7   | 1.45                   | 2.6   | 4.49                   |
| 2.6   | 6.51                   | 5.2   | 9.90                   |
| 2.9   | 2.82                   | 3.3   | 8.93                   |
| 5.2   | 10.45                  | 3.2   | 7.60                   |
| 5.9   | 9.94                   | 4.9   | 3.71                   |
| 3.0   | 4.43                   | 5.5   | 5.47                   |
| 2.4   | 4.75                   | 2.9   | 8.22                   |
| 2.4   | 7.30                   | 2.2   | 7.17                   |
| 0.5   | 3.33                   | 2.3   | 4.35                   |
| 5.0   | 6.76                   |   |                        |

- a. Draw a scatter diagram. Use store size as the independent variable. Does there appear to be a relationship between the two variables. Is it positive or negative?
  - b. Determine the correlation coefficient and the coefficient of determination. Is the relationship strong or weak? Why?
  - c. At the .05 significance level, can we conclude there is a significant positive correlation?
45. **FILE** The manufacturer of Cardio Glide exercise equipment wants to study the relationship between the number of months since the glide was purchased and the time, in hours, the equipment was used last week.

| Person     | Months Owned | Hours Exercised | Person    | Months Owned | Hours Exercised |
|------------|--------------|-----------------|-----------|--------------|-----------------|
| Rupple     | 12           | 4               | Massa     | 2            | 8               |
| Hall       | 2            | 10              | Sass      | 8            | 3               |
| Bennett    | 6            | 8               | Karl      | 4            | 8               |
| Longnecker | 9            | 5               | Malrooney | 10           | 2               |
| Phillips   | 7            | 5               | Veights   | 5            | 5               |

- a. Plot the information on a scatter diagram. Let hours of exercise be the dependent variable. Comment on the graph.
  - b. Determine the correlation coefficient. Interpret.
  - c. At the .01 significance level, can we conclude that there is a negative association between the variables?
46. The following regression equation was computed from a sample of 20 observations:

$$\hat{y} = 15 - 5x$$

- SSE was found to be 100 and SS total was 400.
- a. Determine the standard error of estimate.
  - b. Determine the coefficient of determination.
  - c. Determine the correlation coefficient. (Caution: Watch the sign!)
47. **FILE** City planners believe that larger cities are populated by older residents. To investigate the relationship, data on population and median age in 10 large cities were collected.

| City             | City Population (in millions) | Median Age |
|------------------|-------------------------------|------------|
| Chicago, IL      | 2.833                         | 31.5       |
| Dallas, TX       | 1.233                         | 30.5       |
| Houston, TX      | 2.144                         | 30.9       |
| Los Angeles, CA  | 3.849                         | 31.6       |
| New York, NY     | 8.214                         | 34.2       |
| Philadelphia, PA | 1.448                         | 34.2       |
| Phoenix, AZ      | 1.513                         | 30.7       |
| San Antonio, TX  | 1.297                         | 31.7       |
| San Diego, CA    | 1.257                         | 32.5       |
| San Jose, CA     | 0.930                         | 32.6       |

- a. Plot these data on a scatter diagram with median age as the dependent variable.
- b. Find the correlation coefficient.
- c. A regression analysis was performed and the resulting regression equation is Median Age = 31.4 + 0.272 Population. Interpret the meaning of the slope.
- d. Estimate the median age for a city of 2.5 million people.
- e. Here is a portion of the regression software output. What does it tell you?

| Predictor  | Coef    | SE Coef | T     | P     |
|------------|---------|---------|-------|-------|
| Constant   | 31.3672 | 0.6158  | 50.94 | 0.000 |
| Population | 0.2722  | 0.1901  | 1.43  | 0.190 |

- f. Using the .10 significance level, test the significance of the slope. Interpret the result. Is there a significant relationship between the two variables?

48. **FILE** Emily Smith decides to buy a fuel-efficient used car. Here are several vehicles she is considering, with the estimated cost to purchase and the age of the vehicle.

| Vehicle            | Estimated Cost | Age |
|--------------------|----------------|-----|
| Honda Insight      | \$5,555        | 8   |
| Toyota Prius       | \$17,888       | 3   |
| Toyota Prius       | \$9,963        | 6   |
| Toyota Echo        | \$6,793        | 5   |
| Honda Civic Hybrid | \$10,774       | 5   |
| Honda Civic Hybrid | \$16,310       | 2   |
| Chevrolet Cruz     | \$2,475        | 8   |
| Mazda3             | \$2,808        | 10  |
| Toyota Corolla     | \$7,073        | 9   |
| Acura Integra      | \$8,978        | 8   |
| Scion xB           | \$11,213       | 2   |
| Scion xA           | \$9,463        | 3   |
| Mazda3             | \$15,055       | 2   |
| Mini Cooper        | \$20,705       | 2   |

- Plot these data on a scatter diagram with estimated cost as the dependent variable.
- Find the correlation coefficient.
- A regression analysis was performed and the resulting regression equation is Estimated Cost = 18358 – 1534 Age. Interpret the meaning of the slope.
- Estimate the cost of a five-year-old car.
- Here is a portion of the regression software output. What does it tell you?

| Predictor | Coef    | SE Coef | T     | P     |
|-----------|---------|---------|-------|-------|
| Constant  | 18358   | 1817    | 10.10 | 0.000 |
| Age       | -1533.6 | 306.3   | -5.01 | 0.000 |

- Using the .10 significance level, test the significance of the slope. Interpret the result. Is there a significant relationship between the two variables?
49. **FILE** The National Highway Association is studying the relationship between the number of bidders on a highway project and the winning (lowest) bid for the project. Of particular interest is whether the number of bidders increases or decreases the amount of the winning bid.

| Project | Number of Bidders, $x$ | Winning Bid (\$ millions), $y$ | Project | Number of Bidders, $x$ | Winning Bid (\$ millions), $y$ |
|---------|------------------------|--------------------------------|---------|------------------------|--------------------------------|
| 1       | 9                      | 5.1                            | 9       | 6                      | 10.3                           |
| 2       | 9                      | 8.0                            | 10      | 6                      | 8.0                            |
| 3       | 3                      | 9.7                            | 11      | 4                      | 8.8                            |
| 4       | 10                     | 7.8                            | 12      | 7                      | 9.4                            |
| 5       | 5                      | 7.7                            | 13      | 7                      | 8.6                            |
| 6       | 10                     | 5.5                            | 14      | 7                      | 8.1                            |
| 7       | 7                      | 8.3                            | 15      | 6                      | 7.8                            |
| 8       | 11                     | 5.5                            |         |                        |                                |

- Determine the regression equation. Interpret the equation. Do more bidders tend to increase or decrease the amount of the winning bid?
- Estimate the amount of the winning bid if there were seven bidders.
- A new entrance is to be constructed on the Ohio Turnpike. There are seven bidders on the project. Develop a 95% prediction interval for the winning bid.
- Determine the coefficient of determination. Interpret its value.

50. **FILE** Mr. William Profit is studying companies going public for the first time. He is particularly interested in the relationship between the size of the offering and the price per share. A sample of 15 companies that recently went public revealed the following information.

| Company | Size<br>(\$ millions),<br>$x$ | Price<br>per Share,<br>$y$ | Company | Size<br>(\$ millions),<br>$x$ | Price<br>per Share,<br>$y$ |
|---------|-------------------------------|----------------------------|---------|-------------------------------|----------------------------|
| 1       | 9.0                           | 10.8                       | 9       | 160.7                         | 11.3                       |
| 2       | 94.4                          | 11.3                       | 10      | 96.5                          | 10.6                       |
| 3       | 27.3                          | 11.2                       | 11      | 83.0                          | 10.5                       |
| 4       | 179.2                         | 11.1                       | 12      | 23.5                          | 10.3                       |
| 5       | 71.9                          | 11.1                       | 13      | 58.7                          | 10.7                       |
| 6       | 97.9                          | 11.2                       | 14      | 93.8                          | 11.0                       |
| 7       | 93.5                          | 11.0                       | 15      | 34.4                          | 10.8                       |
| 8       | 70.0                          | 10.7                       |         |                               |                            |

- Determine the regression equation.
  - Conduct a test to determine whether the slope of the regression line is positive.
  - Determine the coefficient of determination. Do you think Mr. Profit should be satisfied with using the size of the offering as the independent variable?
51. **FILE** Bardi Trucking Co., located in Cleveland, Ohio, makes deliveries in the Great Lakes region, the Southeast, and the Northeast. Jim Bardi, the president, is studying the relationship between the distance a shipment must travel and the length of time, in days, it takes the shipment to arrive at its destination. To investigate, Mr. Bardi selected a random sample of 20 shipments made last month. Shipping distance is the independent variable and shipping time is the dependent variable. The results are as follows:

| Shipment | Distance<br>(miles) | Shipping Time<br>(days) | Shipment | Distance<br>(miles) | Shipping Time<br>(days) |
|----------|---------------------|-------------------------|----------|---------------------|-------------------------|
| 1        | 656                 | 5                       | 11       | 862                 | 7                       |
| 2        | 853                 | 14                      | 12       | 679                 | 5                       |
| 3        | 646                 | 6                       | 13       | 835                 | 13                      |
| 4        | 783                 | 11                      | 14       | 607                 | 3                       |
| 5        | 610                 | 8                       | 15       | 665                 | 8                       |
| 6        | 841                 | 10                      | 16       | 647                 | 7                       |
| 7        | 785                 | 9                       | 17       | 685                 | 10                      |
| 8        | 639                 | 9                       | 18       | 720                 | 8                       |
| 9        | 762                 | 10                      | 19       | 652                 | 6                       |
| 10       | 762                 | 9                       | 20       | 828                 | 10                      |

- Draw a scatter diagram. Based on these data, does it appear that there is a relationship between how many miles a shipment has to go and the time it takes to arrive at its destination?
  - Determine the correlation coefficient. Can we conclude that there is a positive correlation between distance and time? Use the .05 significance level.
  - Determine and interpret the coefficient of determination.
  - Determine the standard error of estimate.
  - Would you recommend using the regression equation to predict shipping time? Why or why not?
52. **FILE** Super Markets Inc. is considering expanding into the Scottsdale, Arizona, area. You, as director of planning, must present an analysis of the proposed expansion to the operating committee of the board of directors. As a part of your proposal, you need to include information on the amount people in the region spend per month for grocery items. You would also like to include information on the relationship between the amount spent for grocery items and income. Your assistant gathered the following sample information.

| Household | Amount Spent | Monthly Income |
|-----------|--------------|----------------|
| 1         | \$ 555       | \$4,388        |
| 2         | 489          | 4,558          |
| ⋮         | ⋮            | ⋮              |
| 39        | 1,206        | 9,862          |
| 40        | 1,145        | 9,883          |

- Let the amount spent be the dependent variable and monthly income the independent variable. Create a scatter diagram using a software package.
  - Determine the regression equation. Interpret the slope value.
  - Determine the correlation coefficient. Can you conclude that it is greater than 0?
53. **FILE** Below is information on the price per share and the dividend for a sample of 30 companies.

| Company | Price per Share | Dividend |
|---------|-----------------|----------|
| 1       | \$20.00         | \$ 3.14  |
| 2       | 22.01           | 3.36     |
| ⋮       | ⋮               | ⋮        |
| 29      | 77.91           | 17.65    |
| 30      | 80.00           | 17.36    |

- Calculate the regression equation using selling price based on the annual dividend.
  - Test the significance of the slope.
  - Determine the coefficient of determination. Interpret its value.
  - Determine the correlation coefficient. Can you conclude that it is greater than 0 using the .05 significance level?
54. A highway employee performed a regression analysis of the relationship between the number of construction work-zone fatalities and the number of unemployed people in a state. The regression equation is  $\text{Fatalities} = 12.7 + 0.000114 (\text{Unemp})$ . Some additional output is:

| Predictor            | Coef       | SE Coef    | T     | P     |       |
|----------------------|------------|------------|-------|-------|-------|
| Constant             | 12.726     | 8.115      | 1.57  | 0.134 |       |
| Unemp                | 0.00011386 | 0.00002896 | 3.93  | 0.001 |       |
| Analysis of Variance |            |            |       |       |       |
| Source               | DF         | SS         | MS    | F     | P     |
| Regression           | 1          | 10354      | 10354 | 15.46 | 0.001 |
| Residual Error       | 18         | 12054      | 670   |       |       |
| Total                | 19         | 22408      |       |       |       |

- How many states were in the sample?
  - Determine the standard error of estimate.
  - Determine the coefficient of determination.
  - Determine the correlation coefficient.
  - At the .05 significance level, does the evidence suggest there is a positive association between fatalities and the number unemployed?
55. A regression analysis relating the current market value in dollars to the size in square feet of homes in Greene County, Tennessee, follows. The regression equation is:  $\text{Value} = -37,186 + 65.0 \text{ Size}$ .

| Predictor            | Coef   | SE Coef     | T           | P      |       |
|----------------------|--------|-------------|-------------|--------|-------|
| Constant             | -37186 | 4629        | -8.03       | 0.000  |       |
| Size                 | 64.993 | 3.047       | 21.33       | 0.000  |       |
| Analysis of Variance |        |             |             |        |       |
| Source               | DF     | SS          | MS          | F      | P     |
| Regression           | 1      | 13548662082 | 13548662082 | 454.98 | 0.000 |
| Residual Error       | 33     | 982687392   | 29778406    |        |       |
| Total                | 34     | 14531349474 |             |        |       |

- a. How many homes were in the sample?
  - b. Compute the standard error of estimate.
  - c. Compute the coefficient of determination.
  - d. Compute the correlation coefficient.
  - e. At the .05 significance level, does the evidence suggest a positive association between the market value of homes and the size of the home in square feet?
56. **FILE** The following table shows the mean annual percent return on capital (profitability) and the mean annual percentage sales growth for eight aerospace and defense companies.

| Company             | Profitability | Growth |
|---------------------|---------------|--------|
| Alliant Techsystems | 23.1          | 8.0    |
| Boeing              | 13.2          | 15.6   |
| General Dynamics    | 24.2          | 31.2   |
| Honeywell           | 11.1          | 2.5    |
| L-3 Communications  | 10.1          | 35.4   |
| Northrop Grumman    | 10.8          | 6.0    |
| Rockwell Collins    | 27.3          | 8.7    |
| United Technologies | 20.1          | 3.2    |

- a. Compute the correlation coefficient. Conduct a test of hypothesis to determine if it is reasonable to conclude that the population correlation is greater than zero. Use the .05 significance level.
  - b. Develop the regression equation for profitability based on growth. Can we conclude that the slope of the regression line is negative?
  - c. Use a software package to determine the residual for each observation. Which company has the largest residual?
57. **FILE** The following data show the retail price for 12 randomly selected laptop computers along with their corresponding processor speeds in gigahertz.

| Computer | Speed | Price     | Computer | Speed | Price     |
|----------|-------|-----------|----------|-------|-----------|
| 1        | 2.0   | \$1008.50 | 7        | 2.0   | \$1098.50 |
| 2        | 1.6   | 461.00    | 8        | 1.6   | 693.50    |
| 3        | 1.6   | 532.00    | 9        | 2.0   | 1057.00   |
| 4        | 1.8   | 971.00    | 10       | 1.6   | 1001.00   |
| 5        | 2.0   | 1068.50   | 11       | 1.0   | 468.50    |
| 6        | 1.2   | 506.00    | 12       | 1.4   | 434.50    |

- a. Develop a linear equation that can be used to describe how the price depends on the processor speed.
  - b. Based on your regression equation, is there one machine that seems particularly over- or underpriced?
  - c. Compute the correlation coefficient between the two variables. At the .05 significance level, conduct a test of hypothesis to determine if the population correlation is greater than zero.
58. **FILE** A consumer buying cooperative tested the effective heating area of 20 different electric space heaters with different wattages. Here are the results.

| Heater | Wattage | Area | Heater | Wattage | Area |
|--------|---------|------|--------|---------|------|
| 1      | 1,500   | 205  | 11     | 1,250   | 116  |
| 2      | 750     | 70   | 12     | 500     | 72   |
| 3      | 1,500   | 199  | 13     | 500     | 82   |
| 4      | 1,250   | 151  | 14     | 1,500   | 206  |
| 5      | 1,250   | 181  | 15     | 2,000   | 245  |
| 6      | 1,250   | 217  | 16     | 1,500   | 219  |
| 7      | 1,000   | 94   | 17     | 750     | 63   |
| 8      | 2,000   | 298  | 18     | 1,500   | 200  |
| 9      | 1,000   | 135  | 19     | 1,250   | 151  |
| 10     | 1,500   | 211  | 20     | 500     | 44   |

- a. Compute the correlation between the wattage and heating area. Is there a direct or an indirect relationship?
- b. Conduct a test of hypothesis to determine if it is reasonable that the coefficient is greater than zero. Use the .05 significance level.
- c. Develop the regression equation for effective heating based on wattage.
- d. Which heater looks like the “best buy” based on the size of the residual?
59. **FILE** A dog trainer is exploring the relationship between the size of the dog (weight in pounds) and its daily food consumption (measured in standard cups). Below is the result of a sample of 18 observations.

| Dog | Weight | Consumption | Dog | Weight | Consumption |
|-----|--------|-------------|-----|--------|-------------|
| 1   | 41     | 3           | 10  | 91     | 5           |
| 2   | 148    | 8           | 11  | 109    | 6           |
| 3   | 79     | 5           | 12  | 207    | 10          |
| 4   | 41     | 4           | 13  | 49     | 3           |
| 5   | 85     | 5           | 14  | 113    | 6           |
| 6   | 111    | 6           | 15  | 84     | 5           |
| 7   | 37     | 3           | 16  | 95     | 5           |
| 8   | 111    | 6           | 17  | 57     | 4           |
| 9   | 41     | 3           | 18  | 168    | 9           |

- a. Compute the correlation coefficient. Is it reasonable to conclude that the correlation in the population is greater than zero? Use the .05 significance level.
- b. Develop the regression equation for cups based on the dog's weight. How much does each additional cup change the estimated weight of the dog?
- c. Is one of the dogs a big undereater or overeater?
60. Waterbury Insurance Company wants to study the relationship between the amount of fire damage and the distance between the burning house and the nearest fire station. This information will be used in setting rates for insurance coverage. For a sample of 30 claims for the last year, the director of the actuarial department determined the distance from the fire station ( $x$ ) and the amount of fire damage, in thousands of dollars ( $y$ ). The MegaStat output is reported below.

| ANOVA table       |              |            |            |       |
|-------------------|--------------|------------|------------|-------|
| Source            | SS           | df         | MS         | F     |
| Regression        | 1,864.5782   | 1          | 1,864.5782 | 38.83 |
| Residual          | 1,344.4934   | 28         | 48.0176    |       |
| Total             | 3,209.0716   | 29         |            |       |
| Regression output |              |            |            |       |
| Variables         | Coefficients | Std. Error | t(df = 28) |       |
| Intercept         | 12.3601      | 3.2915     | 3.755      |       |
| Distance-X        | 4.7956       | 0.7696     | 6.231      |       |

Answer the following questions.

- a. Write out the regression equation. Is there a direct or indirect relationship between the distance from the fire station and the amount of fire damage?
- b. How much damage would you estimate for a fire 5 miles from the nearest fire station?
- c. Determine and interpret the coefficient of determination.
- d. Determine the correlation coefficient. Interpret its value. How did you determine the sign of the correlation coefficient?
- e. Conduct a test of hypothesis to determine if there is a significant relationship between the distance from the fire station and the amount of damage. Use the .01 significance level and a two-tailed test.

61. **FILE** TravelAir.com samples domestic airline flights to explore the relationship between airfare and distance. The service would like to know if there is a correlation between airfare and flight distance. If there is a correlation, what percentage of the variation in airfare is accounted for by distance? How much does each additional mile add to the fare? The data follow.

| Origin              | Destination          | Distance | Fare  |
|---------------------|----------------------|----------|-------|
| Detroit, MI         | Myrtle Beach, SC     | 636      | \$109 |
| Baltimore, MD       | Sacramento, CA       | 2,395    | 252   |
| Las Vegas, NV       | Philadelphia, PA     | 2,176    | 221   |
| Sacramento, CA      | Seattle, WA          | 605      | 151   |
| Atlanta, GA         | Orlando, FL          | 403      | 138   |
| Boston, MA          | Miami, FL            | 1,258    | 209   |
| Chicago, IL         | Covington, KY        | 264      | 254   |
| Columbus, OH        | Minneapolis, MN      | 627      | 259   |
| Fort Lauderdale, FL | Los Angeles, CA      | 2,342    | 215   |
| Chicago, IL         | Indianapolis, IN     | 177      | 128   |
| Philadelphia, PA    | San Francisco, CA    | 2,521    | 348   |
| Houston, TX         | Raleigh/Durham, NC   | 1,050    | 224   |
| Houston, TX         | Midland/Odessa, TX   | 441      | 175   |
| Cleveland, OH       | Dallas/Ft. Worth, TX | 1,021    | 256   |
| Baltimore, MD       | Columbus, OH         | 336      | 121   |
| Boston, MA          | Covington, KY        | 752      | 252   |
| Kansas City, MO     | San Diego, CA        | 1,333    | 206   |
| Milwaukee, WI       | Phoenix, AZ          | 1,460    | 167   |
| Portland, OR        | Washington, DC       | 2,350    | 308   |
| Phoenix, AZ         | San Jose, CA         | 621      | 152   |
| Baltimore, MD       | St. Louis, MO        | 737      | 175   |
| Houston, TX         | Orlando, FL          | 853      | 191   |
| Houston, TX         | Seattle, WA          | 1,894    | 231   |
| Burbank, CA         | New York, NY         | 2,465    | 251   |
| Atlanta, GA         | San Diego, CA        | 1,891    | 291   |
| Minneapolis, MN     | New York, NY         | 1,028    | 260   |
| Atlanta, GA         | West Palm Beach, FL  | 545      | 123   |
| Kansas City, MO     | Seattle, WA          | 1,489    | 211   |
| Baltimore, MD       | Portland, ME         | 452      | 139   |
| New Orleans, LA     | Washington, DC       | 969      | 243   |

- Draw a scatter diagram with *Distance* as the independent variable and *Fare* as the dependent variable. Is the relationship direct or indirect?
- Compute the correlation coefficient. At the .05 significance level, is it reasonable to conclude that the correlation coefficient is greater than zero?
- What percentage of the variation in *Fare* is accounted for by *Distance* of a flight?
- Determine the regression equation. How much does each additional mile add to the fare? Estimate the fare for a 1,500-mile flight.
- A traveler is planning to fly from Atlanta to London Heathrow. The distance is 4,218 miles. She wants to use the regression equation to estimate the fare. Explain why it would not be a good idea to estimate the fare for this international flight with the regression equation.

## DATA ANALYTICS

(The data for these exercises are available in Connect.)

62. **FILE** The North Valley Real Estate data report information on homes on the market.
- Let selling price be the dependent variable and size of the home the independent variable. Determine the regression equation. Estimate the selling price for a home with an area of 2,200 square feet. Determine the 95% confidence interval for all 2,200-square-foot homes and the 95% prediction interval for the selling price of a home with 2,200 square feet.



- b. Let days-on-the-market be the dependent variable and price be the independent variable. Determine the regression equation. Estimate the days-on-the-market of a home that is priced at \$300,000. Determine the 95% confidence interval of days-on-the-market for homes with a mean price of \$300,000 and the 95% prediction interval of days-on-the-market for a home priced at \$300,000.
- c. Can you conclude that the independent variables “days on the market” and “selling price” are positively correlated? Are the size of the home and the selling price positively correlated? Use the .05 significance level. Report the  $p$ -value of the test. Summarize your results in a brief report.
63. **FILE** Refer to the Baseball 2016 data, which report information on the 2016 Major League Baseball season. Let attendance be the dependent variable and total team salary be the independent variable. Determine the regression equation and answer the following questions.
- Draw a scatter diagram. From the diagram, does there seem to be a direct relationship between the two variables?
  - What is the expected attendance for a team with a salary of \$100.0 million?
  - If the owners pay an additional \$30 million, how many more people could they expect to attend?
  - At the .05 significance level, can we conclude that the slope of the regression line is positive? Conduct the appropriate test of hypothesis.
  - What percentage of the variation in attendance is accounted for by salary?
  - Determine the correlation between attendance and team batting average and between attendance and team ERA. Which is stronger? Conduct an appropriate test of hypothesis for each set of variables.
64. **FILE** Refer to the Lincolnville School bus data. Develop a regression equation that expresses the relationship between age of the bus and maintenance cost. The age of the bus is the independent variable.
- Draw a scatter diagram. What does this diagram suggest as to the relationship between the two variables? Is it direct or indirect? Does it appear to be strong or weak?
  - Develop a regression equation. How much does an additional year add to the maintenance cost. What is the estimated maintenance cost for a 10-year-old bus?
  - Conduct a test of hypothesis to determine whether the slope of the regression line is greater than zero. Use the .05 significance level. Interpret your findings from parts (a), (b), and (c) in a brief report.

## PRACTICE TEST

### Part 1—Objective

- The first step in correlation analysis is to plot the data with a \_\_\_\_\_.
- The range of the correlation coefficient is between \_\_\_\_\_ and \_\_\_\_\_.
- In studying the relationship between two variables, if the value of one variable decreases with increases in the other variable, the correlation coefficient is \_\_\_\_\_. (less than zero, zero, greater than zero)
- The proportion of variation in the dependent variable that is explained by the variation in the independent variable is measured by the \_\_\_\_\_.
- To test the hypothesis that the correlation coefficient is zero, the test statistic follows the \_\_\_\_\_ distribution.
- The least squares regression line minimizes the sum of the squared differences between the actual and \_\_\_\_\_ values of the dependent variable.
- For a given set of data, the correlation coefficient and the slope of the regression line have the same \_\_\_\_\_. (values, signs, units, squares)
- For a regression analysis, a small standard error of the estimate indicates that the coefficient of determination will be \_\_\_\_\_. (large, small, always 0)
- In regression analysis, confidence and prediction intervals show the \_\_\_\_\_ associated with an estimated value of the dependent variable. (error, association, convergence, sample size)
- A prediction interval is based on an individual value of the \_\_\_\_\_ variable. (dependent, independent, correlated, estimated)

**Part 2—Problems**

1. At the end of each calendar year, employees of the G. G. Green Manufacturing Company can purchase company stock. For a sample of employees, the Director of Human Resources investigated the relationship between the number of years of service with the company and the number of shares of company stock owned. The “number of years of service” is used to estimate the “number of shares of stock.” Use the following output showing the results of the analysis to answer the questions.

| ANOVA Table |              |    |              |       |
|-------------|--------------|----|--------------|-------|
| Source      | SS           | df | MS           | F     |
| Regression  | 152,399.0211 | 1  | 152,399.0211 | 62.67 |
| Residual    | 55,934.1189  | 23 | 2,431.9182   |       |
| Total       | 208,333.1400 | 24 |              |       |

| Regression Output |              |            |             |
|-------------------|--------------|------------|-------------|
| Variables         | Coefficients | Std. Error | t (df = 23) |
| Intercept         | 197.9229     | 34.3047    | 5.770       |
| Years             | 24.9145      | 3.1473     | 7.916       |

- How many employees were included in the study?
- Is “number of shares of stock” or “number of years of service” the dependent variable?
- Write out the regression equation.
- Is the relationship between the two variables direct or indirect?
- Determine the correlation coefficient.
- How many shares would you expect an employee of 10 years to own?
- For each additional year of service, how much does “number of shares owned” change?
- Can we conclude that as years of service increases so do the number of shares of stock owned? Conduct an appropriate test of hypothesis on the slope of the regression line.

# 14

# Multiple Regression Analysis



©PhotoAlto/Alamy

- ▲ **THE MORTGAGE DEPARTMENT** of the Bank of New England is studying data from recent loans. Of particular interest is how such factors as the value of the home being purchased, education level of the head of the household, age of the head of the household, current monthly mortgage payment, and gender of the head of the household relate to the family income. Are the proposed variables effective predictors of the dependent variable family income? (See the example/solution within the Review of Multiple Regression section.)

## LEARNING OBJECTIVES

---

*When you have completed this chapter, you will be able to:*

- LO14-1** Use multiple regression analysis to describe and interpret a relationship between several independent variables and a dependent variable.
- LO14-2** Evaluate how well a multiple regression equation fits the data.
- LO14-3** Test hypotheses about the relationships inferred by a multiple regression model.
- LO14-4** Evaluate the assumptions of multiple regression.
- LO14-5** Use and interpret a qualitative, dummy variable in multiple regression.
- LO14-6** Apply stepwise regression to develop a multiple regression model.
- LO14-7** Apply multiple regression techniques to develop a linear model.

## INTRODUCTION

In Chapter 13, we described the relationship between a pair of interval- or ratio-scaled variables. We began the chapter by studying the correlation coefficient, which measures the strength of the relationship. A coefficient near plus or minus 1.00 (−.88 or .78, for example) indicates a very strong linear relationship, whereas a value near 0 (−.12 or .18, for example) indicates that the relationship is weak. Next we developed a procedure to determine a linear equation to express the relationship between the two variables. We referred to this as a *regression line*. This line describes the relationship between the variables. It also describes the overall pattern of a dependent variable ( $y$ ) to a single independent or explanatory variable ( $x$ ).

In multiple linear correlation and regression, we use additional independent variables (denoted  $x_1, x_2, \dots$ , and so on) that help us better explain or predict the dependent variable ( $y$ ). Almost all of the ideas we saw in simple linear correlation and regression extend to this more general situation. However, the additional independent variables do lead to some new considerations. Multiple regression analysis can be used either as a descriptive or as an inferential technique.

### LO14-1

Use multiple regression analysis to describe and interpret a relationship between several independent variables and a dependent variable.

## MULTIPLE REGRESSION ANALYSIS

The general descriptive form of a multiple linear equation is shown in formula (14–1). We use  $k$  to represent the number of independent variables. So  $k$  can be any positive integer.

### GENERAL MULTIPLE REGRESSION EQUATION

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_kx_k \quad (14-1)$$

where:

$a$  is the  $y$ -intercept, the value of  $\hat{y}$  when all the  $x$ 's are zero.

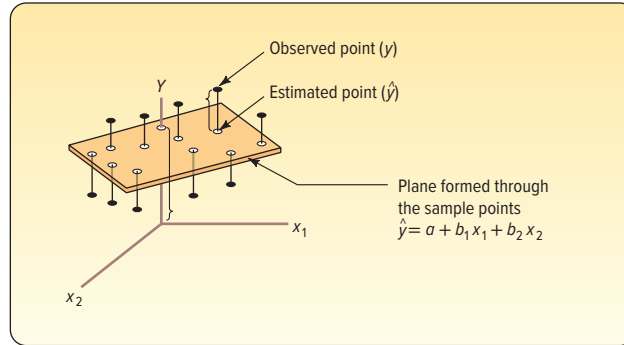
$b_j$  is the amount by which  $\hat{y}$  changes when that particular  $x_j$  increases by one unit, with the values of all other independent variables held constant. The subscript  $j$  is simply a label that helps to identify each independent variable; it is not used in any calculations. Usually the subscript is an integer value between 1 and  $k$ , which is the number of independent variables. However, the subscript can also be a short or abbreviated label. For example, “age” could be used as a subscript to identify the independent variable, age.

In Chapter 13, the regression analysis described and tested the relationship between a dependent variable,  $\hat{y}$ , and a single independent variable,  $x$ . The relationship between  $\hat{y}$  and  $x$  was graphically portrayed by a line. When there are two independent variables, the regression equation is

$$\hat{y} = a + b_1x_1 + b_2x_2$$

Because there are two independent variables, this relationship is graphically portrayed as a plane and is shown in Chart 14–1. The chart shows the residuals as the difference between the actual  $y$  and the fitted  $\hat{y}$  on the plane. If a multiple regression analysis includes more than two independent variables, we cannot use a graph to illustrate the analysis since graphs are limited to three dimensions.

To illustrate the interpretation of the intercept and the two regression coefficients, suppose the selling price of a home is directly related to the number of rooms and *inversely* related to its age. We let  $x_1$  refer to the number of rooms,  $x_2$  to the age of the home in years, and  $y$  to the selling price of the home in thousands of dollars (\$000).



**CHART 14–1** Regression Plane with 10 Sample Points

Suppose the regression equation, calculated using statistical software, is:

$$\hat{y} = 21.2 + 18.7x_1 - 0.25x_2$$

The intercept value of 21.2 indicates the regression equation (plane) intersects the  $y$ -axis at 21.2. This happens when both the number of rooms and the age of the home are zero. We could say that \$21,200 is the average value of a property without a house.

The first regression coefficient, 18.7, indicates that for each increase of one room in the size of a home, the selling price will increase by \$18.7 thousand (\$18,700), regardless of the age of the home. The second regression coefficient,  $-0.25$ , indicates that for each increase of one year in age, the selling price will *decrease* by \$.25 thousand (\$250), regardless of the number of rooms. As an example, a seven-room home that is 30 years old is expected to sell for \$144,600.

$$\hat{y} = 21.2 + 18.7x_1 - 0.25x_2 = 21.2 + 18.7(7) - 0.25(30) = 144.6$$

The values for the coefficients in the multiple linear equation are found by using the method of least squares. Recall from the previous chapter that the least squares method makes the sum of the squared differences between the fitted and actual values of  $y$  as small as possible, that is, the term  $\Sigma(y - \hat{y})^2$  is minimized. The calculations are very tedious, so they are usually performed by a statistical software package.

In the following example, we show a multiple regression analysis using three independent variables employing Excel that produces a standard set of statistics and reports. Statistical software, such as Minitab, MegaStat, and others, provides advanced regression analysis techniques.

### ▶ EXAMPLE

Salsberry Realty sells homes along the East Coast of the United States. One of the questions most frequently asked by prospective buyers is: If we purchase this home, how much can we expect to pay to heat it during the winter? The research department at Salsberry has been asked to develop some guidelines regarding heating costs for single-family homes. Three variables are thought to relate to the heating costs: (1) the mean daily outside temperature, (2) the number of inches of insulation in the attic, and (3) the age in years of the furnace. To investigate, Salsberry's research department selected a random sample of 20 recently sold homes. It determined the cost to heat each home last January, as well as the January outside temperature in the region, the number of inches of insulation in the attic, and the age of the furnace. The sample information is reported in Table 14–1.

**TABLE 14–1** Factors in January Heating Cost for a Sample of 20 Homes

| Home | Heating Cost (\$) | Mean Outside Temperature (°F) | Attic Insulation (inches) | Age of Furnace (years) |
|------|-------------------|-------------------------------|---------------------------|------------------------|
| 1    | \$250             | 35                            | 3                         | 6                      |
| 2    | 360               | 29                            | 4                         | 10                     |
| 3    | 165               | 36                            | 7                         | 3                      |
| 4    | 43                | 60                            | 6                         | 9                      |
| 5    | 92                | 65                            | 5                         | 6                      |
| 6    | 200               | 30                            | 5                         | 5                      |
| 7    | 355               | 10                            | 6                         | 7                      |
| 8    | 290               | 7                             | 10                        | 10                     |
| 9    | 230               | 21                            | 9                         | 11                     |
| 10   | 120               | 55                            | 2                         | 5                      |
| 11   | 73                | 54                            | 12                        | 4                      |
| 12   | 205               | 48                            | 5                         | 1                      |
| 13   | 400               | 20                            | 5                         | 15                     |
| 14   | 320               | 39                            | 4                         | 7                      |
| 15   | 72                | 60                            | 8                         | 6                      |
| 16   | 272               | 20                            | 5                         | 8                      |
| 17   | 94                | 58                            | 7                         | 3                      |
| 18   | 190               | 40                            | 8                         | 11                     |
| 19   | 235               | 27                            | 9                         | 8                      |
| 20   | 139               | 30                            | 7                         | 5                      |

The data in Table 14–1 are available in Excel worksheet format in Connect. The basic instructions for using Excel for these data are in the **Software Commands** in Appendix C.

Determine the multiple regression equation. Which variables are the independent variables? Which variable is the dependent variable? Discuss the regression coefficients. What does it indicate if some coefficients are positive and some coefficients are negative? What is the intercept value? What is the estimated heating cost for a home if the mean outside temperature is 30 degrees, there are 5 inches of insulation in the attic, and the furnace is 10 years old?

**SOLUTION**

We begin the analysis by defining the dependent and independent variables. The dependent variable is the January heating cost. It is represented by  $y$ . There are three independent variables:

- The mean outside temperature in January, represented by  $x_1$ .
- The number of inches of insulation in the attic, represented by  $x_2$ .
- The age in years of the furnace, represented by  $x_3$ .

Given these definitions, the general form of the multiple regression equation follows. The value  $\hat{y}$  is used to estimate the value of  $y$ .

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3$$

Now that we have defined the regression equation, we are ready to use Excel to compute all the statistics needed for the analysis. The output from Excel is shown on the following page.

To use the regression equation to predict the January heating cost, we need to know the values of the regression coefficients:  $b_1$ ,  $b_2$ , and  $b_3$ . These are highlighted in the software reports. The software uses the variable names or labels associated with

each independent variable. The regression equation intercept,  $\alpha$ , is labeled “intercept” in the Excel output.

|    | A    | B    | C     | D   | E | F   | G         | H          | I         | J        | K                     |
|----|------|------|-------|-----|---|---|-----------|------------|-----------|----------|-----------------------|
| 1  | Cost | Temp | Insul | Age |   | SUMMARY OUTPUT                                    |           |            |           |          |                       |
| 2  | 250  | 35   | 3     | 6   |   |   |           |            |           |          |                       |
| 3  | 360  | 29   | 4     | 10  |   | <i>Regression Statistics</i>                      |           |            |           |          |                       |
| 4  | 165  | 36   | 7     | 3   |   | Multiple R  | 0.897     |            |           |          |                       |
| 5  | 43   | 60   | 6     | 9   |   | R Square  | 0.804     |            |           |          |                       |
| 6  | 92   | 65   | 5     | 6   |   | Adjusted R Square                                 | 0.767     |            |           |          |                       |
| 7  | 200  | 30   | 5     | 5   |   | Standard Error                                    | 51.049    |            |           |          |                       |
| 8  | 355  | 10   | 6     | 7   |   | Observations                                      | 20        |            |           |          |                       |
| 9  | 290  | 7    | 10    | 10  |   |   |           |            |           |          |                       |
| 10 | 230  | 21   | 9     | 11  |   | <i>ANOVA</i>                                      |           |            |           |          |                       |
| 11 | 120  | 55   | 2     | 5   |   |   | <i>df</i> | <i>SS</i>  | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
| 12 | 73   | 54   | 12    | 4   |   | Regression  | 3         | 171220.473 | 57073.491 | 21.901   | 0.000                 |
| 13 | 205  | 48   | 5     | 1   |   | Residual  | 16        | 41695.277  | 2605.955  |          |                       |
| 14 | 400  | 20   | 5     | 15  |   | Total   | 19        | 212915.750 |           |          |                       |
| 15 | 320  | 39   | 4     | 7   |   |   |           |            |           |          |                       |
| 16 | 72   | 60   | 8     | 6   |   | <i>Coefficients Standard Error t Stat P-value</i> |           |            |           |          |                       |
| 17 | 272  | 20   | 5     | 8   |   | Intercept   | 427.194   | 59.601     | 7.168     | 0.000    |                       |
| 18 | 94   | 58   | 7     | 3   |   | Temp  | -4.583    | 0.772      | -5.934    | 0.000    |                       |
| 19 | 190  | 40   | 8     | 11  |   | Insul   | -14.831   | 4.754      | -3.119    | 0.007    |                       |
| 20 | 235  | 27   | 9     | 8   |   | Age   | 6.101     | 4.012      | 1.521     | 0.148    |                       |
| 21 | 139  | 30   | 7     | 5   |   |   |           |            |           |          |                       |

Source: Microsoft Excel

In this case, the estimated regression equation is:

$$\hat{y} = 427.194 - 4.583x_1 - 14.831x_2 + 6.101x_3$$

We can now estimate or predict the January heating cost for a home if we know the mean outside temperature, the inches of insulation, and the age of the furnace. For an example home, the mean outside temperature for the month is 30 degrees ( $x_1$ ), there are 5 inches of insulation in the attic ( $x_2$ ), and the furnace is 10 years old ( $x_3$ ). By substituting the values for the independent variables:

$$\hat{y} = 427.194 - 4.583(30) - 14.831(5) + 6.101(10) = 276.56$$

The estimated January heating cost is \$276.56.

The regression coefficients, and their algebraic signs, also provide information about their individual relationships with the January heating cost. The regression coefficient for mean outside temperature is  $-4.583$ . The coefficient is negative and shows an inverse relationship between heating cost and temperature. This is not surprising. As the outside temperature increases, the cost to heat the home decreases. The numeric value of the regression coefficient provides more information. If the outside temperature increases by 1 degree and the other two independent variables remain constant, we can estimate a decrease of \$4.583 in monthly heating cost. So if the mean temperature in Boston is 25 degrees and it is 35 degrees in Philadelphia, all other things being the same (insulation and age of furnace), we expect the heating cost would be \$45.83 less in Philadelphia.

The attic insulation variable also shows an inverse relationship: the more insulation in the attic, the less the cost to heat the home. So the negative sign for this coefficient is logical. For each additional inch of insulation, we expect the cost to heat the home to decline \$14.83 per month, holding the outside temperature and the age of the furnace constant.

The age of the furnace variable shows a direct relationship. With an older furnace, the cost to heat the home increases. Specifically, for each additional year older the furnace is, we expect the cost to increase \$6.10 per month.

**STATISTICS IN ACTION**

Many studies indicate a woman will earn about 70% of what a man would for the same work. Researchers at the University of Michigan Institute for Social Research found that about one-third of the difference can be explained by such social factors as differences in education, seniority, and work interruptions. The remaining two-thirds is not explained by these social factors.

**SELF-REVIEW 14-1**



There are many restaurants in northeastern South Carolina. They serve beach vacationers in the summer, golfers in the fall and spring, and snowbirds in the winter. Bill and Joyce Tuneall manage several restaurants in the North Jersey area and are considering moving to Myrtle Beach, SC, to open a new restaurant. Before making a final decision, they wish to

investigate existing restaurants and what variables seem to be related to profitability. They gather sample information where profit (reported in \$000) is the dependent variable and the independent variables are:

- $x_1$  the number of parking spaces near the restaurant.
- $x_2$  the number of hours the restaurant is open per week.
- $x_3$  the distance from the SkyWheel, a landmark in Myrtle Beach.
- $x_4$  the number of servers employed.
- $x_5$  the number of years the current owner operated the restaurant.

The following is part of the output obtained using statistical software.

| Predictor | Coefficient | SE Coefficient | t       |
|-----------|-------------|----------------|---------|
| Constant  | 2.50        | 1.50           | 1.667   |
| $x_1$     | 3.00        | 1.50           | 2.000   |
| $x_2$     | 4.00        | 3.00           | 1.333   |
| $x_3$     | -3.00       | 0.20           | -15.000 |
| $x_4$     | 0.20        | 0.05           | 4.000   |
| $x_5$     | 1.00        | 1.50           | 0.667   |

- (a) What is the amount of profit for a restaurant with 40 parking spaces that is open 72 hours per week, is 10 miles from the SkyWheel, has 20 servers, and has been operated by the current owner for 5 years?
- (b) Interpret the values of  $b_2$  and  $b_3$  in the multiple regression equation.

## EXERCISES

1. The director of marketing at Reeves Wholesale Products is studying monthly sales. Three independent variables were selected as estimators of sales: regional population, per capita income, and regional unemployment rate. The regression equation was computed to be (in dollars):

$$\hat{y} = 64,100 + 0.394x_1 + 9.6x_2 - 11,600x_3$$

- a. What is the full name of the equation?
  - b. Interpret the number 64,100.
  - c. What are the estimated monthly sales for a particular region with a population of 796,000, per capita income of \$6,940, and an unemployment rate of 6.0%?
2. Thompson Photo Works purchased several new, highly sophisticated processing machines. The production department needed some guidance with respect to qualifications needed by an operator. Is age a factor? Is the length of service as an operator (in years) important? In order to explore further the factors needed to estimate performance on the new processing machines, four variables were listed:

- $x_1$  = Length of time an employee was in the industry
- $x_2$  = Mechanical aptitude test score
- $x_3$  = Prior on-the-job rating
- $x_4$  = Age

Performance on the new machine is designated  $y$ .

Thirty employees were selected at random. Data were collected for each, and their performances on the new machines were recorded. A few results are:

| Name          | Performance on New Machine, $y$ | Length of Time in Industry, $x_1$ | Mechanical Aptitude Score, $x_2$ | Prior on-the-Job Performance, $x_3$ | Age, $x_4$ |
|---------------|---------------------------------|-----------------------------------|----------------------------------|-------------------------------------|------------|
| Mike Miraglia | 112                             | 12                                | 312                              | 121                                 | 52         |
| Sue Trythall  | 113                             | 2                                 | 380                              | 123                                 | 27         |



The equation is:

$$\hat{y} = 11.6 + 0.4x_1 + 0.286x_2 + 0.112x_3 + 0.002x_4$$

- a. What is this equation called?
  - b. How many dependent variables are there? Independent variables?
  - c. What is the number 0.286 called?
  - d. As age increases by one year, how much does estimated performance on the new machine increase?
  - e. Carl Knox applied for a job at Photo Works. He has been in the business for 6 years and scored 280 on the mechanical aptitude test. Carl's prior on-the-job performance rating is 97, and he is 35 years old. Estimate Carl's performance on the new machine.
3. A consulting group was hired by the human resources department at General Mills Inc. to survey company employees regarding their degree of satisfaction with their quality of life. A special index, called the index of satisfaction, was used to measure satisfaction. Six factors were studied, namely, age at the time of first marriage ( $x_1$ ), annual income ( $x_2$ ), number of children living ( $x_3$ ), value of all assets ( $x_4$ ), status of health in the form of an index ( $x_5$ ), and the average number of social activities per week—such as bowling and dancing ( $x_6$ ). Suppose the multiple regression equation is:

$$\hat{y} = 16.24 + 0.017x_1 + 0.0028x_2 + 42x_3 + 0.0012x_4 + 0.19x_5 + 26.8x_6$$

- a. What is the estimated index of satisfaction for a person who first married at 18, has an annual income of \$26,500, has three children living, has assets of \$156,000, has an index of health status of 141, and has 2.5 social activities a week on the average?
  - b. Which would add more to satisfaction, an additional income of \$10,000 a year or two more social activities a week?
4. Cellulon, a manufacturer of home insulation, wants to develop guidelines for builders and consumers on how the thickness of the insulation in the attic of a home and the outdoor temperature affect natural gas consumption. In the laboratory, it varied the insulation thickness and temperature. A few of the findings are:

| Monthly Natural Gas Consumption (cubic feet),<br>$y$ | Thickness of Insulation (inches),<br>$x_1$ | Outdoor Temperature (°F),<br>$x_2$ |
|--|--|------------------------------------|
| 30.3   | 6  | 40                                 |
| 26.9   | 12   | 40                                 |
| 22.1   | 8  | 49                                 |

On the basis of the sample results, the regression equation is:

$$\hat{y} = 62.65 - 1.86x_1 - 0.52x_2$$

- a. How much natural gas can homeowners expect to use per month if they install 6 inches of insulation and the outdoor temperature is 40 degrees F?
- b. What effect would installing 7 inches of insulation instead of 6 have on the monthly natural gas consumption (assuming the outdoor temperature remains at 40 degrees F)?
- c. Why are the regression coefficients  $b_1$  and  $b_2$  negative? Is this logical?

**LO14-2**

Evaluate how well a multiple regression equation fits the data.

## EVALUATING A MULTIPLE REGRESSION EQUATION

Many statistics and statistical methods are used to evaluate the relationship between a dependent variable and more than one independent variable. Our first step was to write the relationship in terms of a multiple regression equation. The next step follows on the concepts presented in Chapter 13 by using the information in an ANOVA table to evaluate how well the equation fits the data.

### The ANOVA Table

As in Chapter 13, the statistical analysis of a multiple regression equation is summarized in an ANOVA table. To review, the total variation of the dependent variable,  $y$ , is divided into two components: (1) *regression*, or the variation of  $y$  explained by all the independent variables, and (2) *the error or residual*, or unexplained variation of  $y$ . These two categories are identified in the first column of an ANOVA table below. The column headed “ $df$ ” refers to the degrees of freedom associated with each category. The total number of degrees of freedom is  $n - 1$ . The number of degrees of freedom in the regression is equal to the number of independent variables in the multiple regression equation. We call the regression degrees of freedom  $k$ . The number of degrees of freedom associated with the error term is equal to the total degrees of freedom,  $n - 1$ , minus the regression degrees of freedom,  $k$ . So, the residual or error degrees of freedom is  $(n - 1) - k$ , and is the same as  $n - (k + 1)$ .

| Source            | $df$          | SS       | MS                        | F         |
|-------------------|---------------|----------|---------------------------|-----------|
| Regression        | $k$           | SSR      | $MSR = SSR/k$             | $MSR/MSE$ |
| Residual or error | $n - (k + 1)$ | SSE      | $MSE = SSE/[n - (k + 1)]$ |           |
| Total             | $n - 1$       | SS total |                           |           |

In the ANOVA table, the column headed “SS” lists the sum of squares for each source of variation: regression, residual or error, and total. The sum of squares is the amount of variation attributable to each source.

The total variation of the dependent variable,  $y$ , is summarized in “SS total.” You should note that this is simply the numerator of the usual formula to calculate any variation—in other words, the sum of the squared deviations from the mean. It is computed as:

$$\text{Total Sum of Squares} = \text{SS total} = \sum(y - \bar{y})^2$$

As we have seen, the total sum of squares is the sum of the regression and residual sum of squares. The regression sum of squares is the sum of the squared differences between the estimated or predicted values,  $\hat{y}$ , and the overall mean of  $y$ . The regression sum of squares is found by:

$$\text{Regression Sum of Squares} = \text{SSR} = \sum(\hat{y} - \bar{y})^2$$

The residual sum of squares is the sum of the squared differences between the observed values of the dependent variable,  $y$ , and their corresponding estimated or predicted values,  $\hat{y}$ . Notice that this difference is the error of estimating or predicting the dependent variable with the multiple regression equation. It is calculated as:

$$\text{Residual or Error Sum of Squares} = \text{SSE} = \sum(y - \hat{y})^2$$

We will use the ANOVA table information from the previous example to evaluate the regression equation to estimate January heating costs.

|    | A    | B    | C     | D   | G | H                     | I            | J              | K         | L       | M              |
|----|------|------|-------|-----|---|-----------------------|--------------|----------------|-----------|---------|----------------|
| 1  | Cost | Temp | Insul | Age |   | SUMMARY OUTPUT        |              |                |           |         |                |
| 2  | 250  | 35   | 3     | 6   |   |                       |              |                |           |         |                |
| 3  | 360  | 29   | 4     | 10  |   | Regression Statistics |              |                |           |         |                |
| 4  | 165  | 36   | 7     | 3   |   | Multiple R            | 0.897        |                |           |         |                |
| 5  | 43   | 60   | 6     | 9   |   | R Square              | 0.804        |                |           |         |                |
| 6  | 92   | 65   | 5     | 6   |   | Adjusted R Square     | 0.767        |                |           |         |                |
| 7  | 200  | 30   | 5     | 5   |   | Standard Error        | 51.049       |                |           |         |                |
| 8  | 355  | 10   | 6     | 7   |   | Observations          | 20           |                |           |         |                |
| 9  | 290  | 7    | 10    | 10  |   | ANOVA                 |              |                |           |         |                |
| 10 | 230  | 21   | 9     | 11  |   |                       | df           | SS             | MS        | F       | Significance F |
| 11 | 120  | 55   | 2     | 5   |   | Regression            | 3            | 171220.473     | 57073.491 | 21.901  | 0.000          |
| 12 | 73   | 54   | 12    | 4   |   | Residual              | 16           | 41695.277      | 2605.955  |         |                |
| 13 | 205  | 48   | 5     | 1   |   | Total                 | 19           | 212915.750     |           |         |                |
| 14 | 400  | 20   | 5     | 15  |   |                       |              |                |           |         |                |
| 15 | 320  | 39   | 4     | 7   |   |                       |              |                |           |         |                |
| 16 | 72   | 60   | 8     | 6   |   |                       | Coefficients | Standard Error | t Stat    | P-value |                |
| 17 | 272  | 20   | 5     | 8   |   | Intercept             | 427.194      | 59.601         | 7.168     | 0.000   |                |
| 18 | 94   | 58   | 7     | 3   |   | Temp                  | -4.583       | 0.772          | -5.934    | 0.000   |                |
| 19 | 190  | 40   | 8     | 11  |   | Insul                 | -14.831      | 4.754          | -3.119    | 0.007   |                |
| 20 | 235  | 27   | 9     | 8   |   | Age                   | 6.101        | 4.012          | 1.521     | 0.148   |                |

Source: Microsoft Excel

## Multiple Standard Error of Estimate

We begin with the **multiple standard error of estimate**. Recall that the standard error of estimate is comparable to the standard deviation. To explain the details of the standard error of estimate, refer to the first sampled home in row 2 in the Excel spreadsheet above. The actual heating cost for the first observation,  $y$ , is \$250; the outside temperature,  $x_1$ , is 35 degrees; the depth of insulation,  $x_2$ , is 3 inches; and the age of the furnace,  $x_3$ , is 6 years. Using the regression equation developed in the previous section, the estimated heating cost for this home is:

$$\begin{aligned}\hat{y} &= 427.194 - 4.583x_1 - 14.831x_2 + 6.101x_3 \\ &= 427.194 - 4.583(35) - 14.831(3) + 6.101(6) \\ &= 258.90\end{aligned}$$

So we would estimate that a home with a mean January outside temperature of 35 degrees, 3 inches of insulation, and a 6-year-old furnace would cost \$258.90 to heat. The actual heating cost was \$250, so the residual—which is the difference between the actual value and the estimated value—is  $y - \hat{y} = 250 - 258.90 = -8.90$ . This difference of \$8.90 is the random or unexplained error for the first home sampled. Our next step is to square this difference—that is, find  $(y - \hat{y})^2 = (250 - 258.90)^2 = (-8.90)^2 = 79.21$ .

If we repeat this calculation for the other 19 observations and sum all 20 squared differences, the total will be the residual or error sum of squares from the ANOVA table. Using this information, we can calculate the multiple standard error of the estimate as:

### MULTIPLE STANDARD ERROR OF ESTIMATE

$$S_{y,123\dots k} = \sqrt{\frac{\sum(y - \hat{y})^2}{n - (k + 1)}} = \sqrt{\frac{\text{SSE}}{n - (k + 1)}} \quad (14-2)$$

where:

$y$  is the actual observation.

$\hat{y}$  is the estimated value computed from the regression equation.

$n$  is the number of observations in the sample.

$k$  is the number of independent variables.

SSE is the residual sum of squares from an ANOVA table.

There is more information in the ANOVA table that can be used to compute the multiple standard error of estimate. The column headed “MS” reports the mean squares for the regression and residual variation. These values are calculated as the sum of squares divided by the corresponding degrees of freedom. The multiple standard error

of estimate is equal to the square root of the residual MS, which is also called the mean square error or the MSE.

$$s_{y,123\dots k} = \sqrt{MSE} = \sqrt{2605.995} = \$51.05$$

How do we interpret the standard error of estimate of 51.05? It is the typical “error” when we use this equation to predict the cost. First, the units are the same as the dependent variable, so the standard error is in dollars, \$51.05. Second, we expect the residuals to be approximately normally distributed, so about 68% of the residuals will be within  $\pm\$51.05$  and about 95% within  $\pm 2(51.05)$ , or  $\pm\$102.10$ . As before with similar measures of dispersion, such as the standard error of estimate in Chapter 13, a smaller multiple standard error indicates a better or more effective predictive equation.

### Coefficient of Multiple Determination

Next, let’s look at the coefficient of multiple determination. Recall from the previous chapter the coefficient of determination is defined as the percent of variation in the dependent variable explained, or accounted for, by the independent variable. In the multiple regression case, we extend this definition as follows.

**COEFFICIENT OF MULTIPLE DETERMINATION** The percent of variation in the dependent variable,  $y$ , explained by the set of independent variables,  $X_1, X_2, X_3, \dots, X_k$ .

The characteristics of the coefficient of multiple determination are:

1. **It is symbolized by a capital  $R$  squared.** In other words, it is written as  $R^2$  because it is calculated as the square of a correlation coefficient.
2. **It can range from 0 to 1.** A value near 0 indicates little association between the set of independent variables and the dependent variable. A value near 1 means a strong association.
3. **It cannot assume negative values.** Any number that is squared or raised to the second power cannot be negative.
4. **It is easy to interpret.** Because  $R^2$  is a value between 0 and 1, it is easy to interpret, compare, and understand.

We can calculate the coefficient of determination from the information found in the ANOVA table. We look in the sum of squares column, which is labeled SS in the Excel output, and use the regression sum of squares, SSR, then divide by the total sum of squares, SS total.

**COEFFICIENT OF MULTIPLE DETERMINATION**  $R^2 = \frac{SSR}{SS \text{ total}} \quad (14-3)$

We can use the regression and the total sum of squares from the ANOVA table highlighted in the Excel output appearing earlier in this section and compute the coefficient of determination.

$$R^2 = \frac{SSR}{SS \text{ total}} = \frac{171,220.473}{212,915.750} = .804$$

How do we interpret this value? We conclude that the independent variables (outside temperature, amount of insulation, and age of furnace) explain, or account for, 80.4% of the variation in heating cost. To put it another way, 19.6% of the variation is due to other sources, such as random error or variables not included in the analysis. Using the ANOVA table, 19.6% is the error sum of squares divided by the

total sum of squares. Knowing that the  $SSR + SSE = SS$  total, the following relationship is true.

$$1 - R^2 = 1 - \frac{SSR}{SS \text{ total}} = \frac{SSE}{SS \text{ total}} = \frac{41,695.277}{212,915.750} = .196$$

## Adjusted Coefficient of Determination

The coefficient of determination tends to increase as more independent variables are added to the multiple regression model. Each new independent variable causes the predictions to be more accurate. That, in turn, makes SSE smaller and SSR larger. Hence,  $R^2$  increases only because the total number of independent variables increases and not because the added independent variable is a good predictor of the dependent variable. In fact, if the number of variables,  $k$ , and the sample size,  $n$ , are equal, the coefficient of determination is 1.0. In practice, this situation is rare and would also be ethically questionable. To balance the effect that the number of independent variables has on the coefficient of multiple determination, statistical software packages use an *adjusted* coefficient of multiple determination.

### ADJUSTED COEFFICIENT OF DETERMINATION

$$R_{\text{adj}}^2 = 1 - \frac{\frac{SSE}{n - (k + 1)}}{\frac{SS_{\text{total}}}{n - 1}} \quad (14-4)$$

The error and total sum of squares are divided by their degrees of freedom. Notice especially the degrees of freedom for the error sum of squares include  $k$ , the number of independent variables. For the cost of heating example, the adjusted coefficient of determination is:

$$R_{\text{adj}}^2 = 1 - \frac{\frac{41,695.277}{20 - (3 + 1)}}{\frac{212,915.750}{20 - 1}} = 1 - \frac{2,605.955}{11,206.092} = 1 - .233 = .767$$

If we compare the  $R^2$  (0.80) to the adjusted  $R^2$  (0.767), the difference in this case is small.

## SELF-REVIEW 14-2



Refer to Self-Review 14-1 on the subject of restaurants in Myrtle Beach. The ANOVA portion of the regression output is presented below.

| Analysis of Variance |    |     |    |
|----------------------|----|-----|----|
| Source               | DF | SS  | MS |
| Regression           | 5  | 100 | 20 |
| Residual Error       | 20 | 40  | 2  |
| Total                | 25 | 140 |    |

- How large was the sample?
- How many independent variables are there?
- How many dependent variables are there?
- Compute the standard error of estimate. About 95% of the residuals will be between what two values?
- Determine the coefficient of multiple determination. Interpret this value.
- Find the coefficient of multiple determination, adjusted for the degrees of freedom.

## EXERCISES

5. Consider the ANOVA table that follows.

| Analysis of Variance |    |         |        |      |       |
|----------------------|----|---------|--------|------|-------|
| Source               | DF | SS      | MS     | F    | P     |
| Regression           | 2  | 77.907  | 38.954 | 4.14 | 0.021 |
| Residual Error       | 62 | 583.693 | 9.414  |      |       |
| Total                | 64 | 661.600 |        |      |       |

- Determine the standard error of estimate. About 95% of the residuals will be between what two values?
- Determine the coefficient of multiple determination. Interpret this value.
- Determine the coefficient of multiple determination, adjusted for the degrees of freedom.

6. Consider the ANOVA table that follows.

| Analysis of Variance |    |         |        |       |
|----------------------|----|---------|--------|-------|
| Source               | DF | SS      | MS     | F     |
| Regression           | 5  | 3710.00 | 742.00 | 12.89 |
| Residual Error       | 46 | 2647.38 | 57.55  |       |
| Total                | 51 | 6357.38 |        |       |

- Determine the standard error of estimate. About 95% of the residuals will be between what two values?
- Determine the coefficient of multiple determination. Interpret this value.
- Determine the coefficient of multiple determination, adjusted for the degrees of freedom.

### LO14-3

Test hypotheses about the relationships inferred by a multiple regression model.

## INFERENCES IN MULTIPLE LINEAR REGRESSION

Thus far, multiple regression analysis has been viewed only as a way to describe the relationship between a dependent variable and several independent variables. However, the least squares method also has the ability to draw inferences or generalizations about the relationship for an entire population. Recall that when you create confidence intervals or perform hypothesis tests as a part of inferential statistics, you view the data as a random sample taken from some population.

In the multiple regression setting, we assume there is an unknown population regression equation that relates the dependent variable to the  $k$  independent variables. This is sometimes called a **model** of the relationship. In symbols we write:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

This equation is analogous to formula (14–1) except the coefficients are now reported as Greek letters. We use the Greek letters to denote *population parameters*. Then under a certain set of assumptions, which will be discussed shortly, the computed values of  $a$  and  $b_i$  are sample statistics. These sample statistics are point estimates of the corresponding population parameters  $\alpha$  and  $\beta_i$ . For example, the sample regression coefficient  $b_2$  is a point estimate of the population parameter  $\beta_2$ . The sampling distribution of these point estimates follows the normal probability distribution. These sampling distributions are each centered at their respective parameter values. To put it another way, the means of the sampling distributions are equal to the parameter values to be estimated. Thus, by using the properties of the sampling distributions of these statistics, inferences about the population parameters are possible.

### Global Test: Testing the Multiple Regression Model

We can test the ability of the independent variables  $X_1, X_2, \dots, X_k$  to explain the behavior of the dependent variable  $Y$ . To put this in question form: Can the dependent variable

be estimated without relying on the independent variables? The test used is referred to as the **global test**. Basically, it investigates whether it is possible that all the independent variables have zero regression coefficients.

**GLOBAL TEST** A test used to determine if any of the set of independent variables has regression coefficients different from zero.

To relate this question to the heating cost example, we will test whether the three independent variables (amount of insulation in the attic, mean daily outside temperature, and age of furnace) effectively estimate home heating costs. In testing the hypothesis, we first state the null hypothesis and the alternate hypothesis in terms of the three population parameters,  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$ . Recall that  $b_1$ ,  $b_2$ , and  $b_3$  are sample regression coefficients and are not used in the hypothesis statements. In the null hypothesis, we test whether the regression coefficients in the population are all zero. The null hypothesis is:

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

The alternate hypothesis is:

$$H_1: \text{Not all the } \beta_i\text{'s are 0.}$$

If the hypothesis test fails to reject the null hypothesis, it implies the regression coefficients are all zero and, logically, are of no value in estimating the dependent variable (heating cost). Should that be the case, we would have to search for some other independent variables—or take a different approach—to predict home heating costs.

To test the null hypothesis that the multiple regression coefficients are all zero, we employ the  $F$  distribution introduced in Chapter 12. We will use the .05 level of significance. Recall these characteristics of the  $F$  distribution:

1. **There is a family of  $F$  distributions.** Each time the degrees of freedom in either the numerator or the denominator change, a new  $F$  distribution is created.
2. **The  $F$  distribution cannot be negative.** The smallest possible value is 0.
3. **It is a continuous distribution.** The distribution can assume an infinite number of values between 0 and positive infinity.
4. **It is positively skewed.** The long tail of the distribution is to the right-hand side. As the number of degrees of freedom increases in both the numerator and the denominator, the distribution approaches the normal probability distribution. That is, the distribution will move toward a symmetric distribution.
5. **It is asymptotic.** As the values of  $X$  increase, the  $F$  curve will approach the horizontal axis, but will never touch it.

The  $F$ -statistic to test the global hypothesis follows. As in Chapter 12, it is the ratio of two variances. In this case, the numerator is the regression sum of squares divided by its degrees of freedom,  $k$ . The denominator is the residual sum of squares divided by its degrees of freedom,  $n - (k + 1)$ . The formula follows.

**GLOBAL TEST** 
$$F = \frac{SSR/k}{SSE/[n - (k + 1)]} \quad (14-5)$$

Using the values from the ANOVA table on page 422, the  $F$ -statistic is

$$F = \frac{SSR/k}{SSE/[n - (k + 1)]} = \frac{171,220.473/3}{41,695.277/[20 - (3 + 1)]} = 21.90$$

Remember that the  $F$ -statistic tests the basic null hypothesis that two variances or, in this case, two mean squares are equal. In our global multiple regression hypothesis test, we will reject the null hypothesis,  $H_0$ , that all regression coefficients are zero when the regression mean square is larger in comparison to the residual mean square. If this

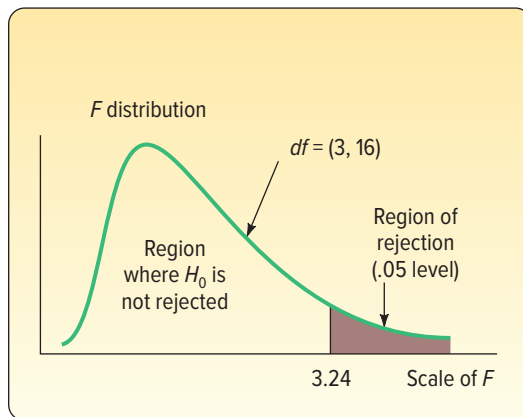
is true, the  $F$ -statistic will be relatively large and in the far-right tail of the  $F$  distribution, and the  $p$ -value will be small, that is, less than our choice of significance level of .05. Thus, we will reject the null hypothesis.

As with other hypothesis-testing methods, the decision rule can be based on either of two methods: (1) comparing the test statistic to a critical value or (2) calculating a  $p$ -value based on the test statistic and comparing the  $p$ -value to the significance level. The critical value method using the  $F$ -statistic requires three pieces of information: (1) the numerator degrees of freedom, (2) the denominator degrees of freedom, and (3) the significance level. The degrees of freedom for the numerator and the denominator are reported in the Excel ANOVA table that follows. The ANOVA output is highlighted in light green. The top number in the column marked “ $df$ ” is 3, indicating there are 3 degrees of freedom in the numerator. This value corresponds to the number of independent variables. The middle number in the “ $df$ ” column (16) indicates that there are 16 degrees of freedom in the denominator. The number 16 is found by  $n - (k + 1) = 20 - (3 + 1) = 16$ .

|    | A    | B    | C     | D   | G | H                     | I              | J          | K         | L              | M     |
|----|------|------|-------|-----|---|-----------------------|----------------|------------|-----------|----------------|-------|
| 1  | Cost | Temp | Insul | Age |   | SUMMARY OUTPUT        |                |            |           |                |       |
| 2  | 250  | 35   | 3     | 6   |   |                       |                |            |           |                |       |
| 3  | 360  | 29   | 4     | 10  |   | Regression Statistics |                |            |           |                |       |
| 4  | 165  | 36   | 7     | 3   |   | Multiple R            | 0.897          |            |           |                |       |
| 5  | 43   | 60   | 6     | 9   |   | R Square              | 0.804          |            |           |                |       |
| 6  | 92   | 65   | 5     | 6   |   | Adjusted R Square     | 0.767          |            |           |                |       |
| 7  | 200  | 30   | 5     | 5   |   | Standard Error        | 51.049         |            |           |                |       |
| 8  | 355  | 10   | 6     | 7   |   | Observations          | 20             |            |           |                |       |
| 9  | 290  | 7    | 10    | 10  |   |                       |                |            |           |                |       |
| 10 | 230  | 21   | 9     | 11  |   | ANOVA                 |                |            |           |                |       |
| 11 | 120  | 55   | 2     | 5   |   | $df$                  | SS             | MS         | F         | Significance F |       |
| 12 | 73   | 54   | 12    | 4   |   | Regression            | 3              | 171220.473 | 57073.491 | 21.901         | 0.000 |
| 13 | 205  | 48   | 5     | 1   |   | Residual              | 16             | 41695.277  | 2605.955  |                |       |
| 14 | 400  | 20   | 5     | 15  |   | Total                 | 19             | 212915.750 |           |                |       |
| 15 | 320  | 39   | 4     | 7   |   |                       |                |            |           |                |       |
| 16 | 72   | 60   | 8     | 6   |   | Coefficients          | Standard Error | t Stat     | P-value   |                |       |
| 17 | 272  | 20   | 5     | 8   |   | Intercept             | 427.194        | 59.601     | 7.168     | 0.000          |       |
| 18 | 94   | 58   | 7     | 3   |   | Temp                  | -4.583         | 0.772      | -5.934    | 0.000          |       |
| 19 | 190  | 40   | 8     | 11  |   | Insul                 | -14.831        | 4.754      | -3.119    | 0.007          |       |
| 20 | 235  | 27   | 9     | 8   |   | Age                   | 6.101          | 4.012      | 1.521     | 0.148          |       |

Source: Microsoft Excel

The critical value of  $F$  is found in Appendix B.6A. Using the table for the .05 significance level, move horizontally to 3 degrees of freedom in the numerator, then down to 16 degrees of freedom in the denominator, and read the critical value. It is 3.24. The region where  $H_0$  is not rejected and the region where  $H_0$  is rejected are shown in the following diagram.



Continuing with the global test, the decision rule is: Do not reject the null hypothesis,  $H_0$ , that all the regression coefficients are 0 if the computed value of  $F$  is less than or equal to 3.24. If the computed  $F$  is greater than 3.24, reject  $H_0$  and accept the alternate hypothesis,  $H_1$ .

The computed value of  $F$  is 21.90, which is in the rejection region. The null hypothesis that all the multiple regression coefficients are zero is therefore rejected. This means that at least one of the independent variables has the ability to explain the variation in the dependent variable (heating cost). We expected this decision. Logically, the



outside temperature, the amount of insulation, or the age of the furnace has a great bearing on heating costs. The global test assures us that they do.

Testing the null hypothesis can also be based on a  $p$ -value, which is reported in the statistical software output for all hypothesis tests. In the case of the  $F$ -statistic, the  $p$ -value is defined as the probability of observing an  $F$ -value as large or larger than the  $F$  test statistic, assuming the null hypothesis is true. If the  $p$ -value is less than our selected significance level, then we decide to reject the null hypothesis. The ANOVA shows the  $F$ -statistic's  $p$ -value is equal to 0.000. It is clearly less than our significance level of .05. Therefore, we decide to reject the global null hypothesis and conclude that at least one of the regression coefficients is not equal to zero.

## Evaluating Individual Regression Coefficients

So far we have shown that at least one, but not necessarily all, of the regression coefficients is not equal to zero and thus useful for predictions. The next step is to test the independent variables *individually* to determine which regression coefficients may be 0 and which are not.

Why is it important to know if any of the  $\beta_i$ 's equal 0? If a  $\beta$  could equal 0, it implies that this particular independent variable is of no value in explaining any variation in the dependent value. If there are coefficients for which  $H_0$  cannot be rejected, we may want to eliminate them from the regression equation.

Our strategy is to use three sets of hypotheses: one for temperature, one for insulation, and one for age of the furnace.

|                       |                       |                       |
|-----------------------|-----------------------|-----------------------|
| For temperature:      | For insulation:       | For furnace age:      |
| $H_0: \beta_1 = 0$    | $H_0: \beta_2 = 0$    | $H_0: \beta_3 = 0$    |
| $H_1: \beta_1 \neq 0$ | $H_1: \beta_2 \neq 0$ | $H_1: \beta_3 \neq 0$ |

We will test the hypotheses at the .05 level. Note that these are two-tailed tests.

The test statistic follows Student's  $t$  distribution with  $n - (k + 1)$  degrees of freedom. The number of sample observations is  $n$ . There are 20 homes in the study, so  $n = 20$ . The number of independent variables is  $k$ , which is 3. Thus, there are  $n - (k + 1) = 20 - (3 + 1) = 16$  degrees of freedom.

The critical value for  $t$  is in Appendix B.5. For a two-tailed test with 16 degrees of freedom using the .05 significance level,  $H_0$  is rejected if  $t$  is less than  $-2.120$  or greater than  $2.120$ .

Refer to the Excel output earlier in this section. The column highlighted in orange, headed Coefficients, shows the values for the multiple regression equation:

$$\hat{y} = 427.194 - 4.583x_1 - 14.831x_2 + 6.101x_3$$

Interpreting the term  $-4.583x_1$  in the equation: For each degree increase in temperature, we predict that heating cost will decrease \$4.58, holding the insulation and age of the furnace variables constant.

The column in the Excel output labeled "Standard Error" shows the standard error of the sample regression coefficients. Recall that Salsberry Realty selected a sample of 20 homes along the East Coast of the United States. If Salsberry Realty selected a second random sample and computed the regression coefficients for that sample, the values would not be exactly the same. If the sampling process was repeated many times, we could construct a sampling distribution for each of these regression coefficients. The column labeled "Standard Error" estimates the variability for each of these regression coefficients. The sampling distributions of the coefficients follow the  $t$  distribution with  $n - (k + 1)$  degrees of freedom. Hence, we are able to test the independent variables individually to determine whether the regression coefficients differ from zero. The formula is:

### TESTING INDIVIDUAL REGRESSION COEFFICIENTS

$$t = \frac{b_i - 0}{s_{b_i}}$$

(14-6)

The  $b_i$  refers to any one of the regression coefficients, and  $s_{b_i}$  refers to the standard deviation of that distribution of the regression coefficient. We include 0 in the equation because the null hypothesis is  $\beta_i = 0$ .

To illustrate this formula, refer to the test of the regression coefficient for the independent variable temperature. From the output earlier in this section, the regression coefficient for temperature is  $-4.583$ . The standard deviation of the sampling distribution of the regression coefficient for the independent variable temperature is  $0.772$ . Inserting these values in formula (14–6):

$$t = \frac{b_1 - 0}{s_{b_1}} = \frac{-4.583 - 0}{0.772} = -5.937$$

The computed value of  $t$  is  $-5.937$  for temperature (the small difference between the computed value and that shown on the Excel output is due to rounding) and  $-3.119$  for insulation. Both of these  $t$ -values are in the rejection region to the left of  $-2.120$ . Thus, we conclude that the regression coefficients for the temperature and insulation variables are *not* zero. The computed  $t$  for the age of the furnace is  $1.521$ , so we conclude that the coefficient could equal 0. The independent variable age of the furnace is not a significant predictor of heating cost. The results of these hypothesis tests indicate that the analysis should focus on temperature and insulation as predictors of heating cost.

We can also use  $p$ -values to test the individual regression coefficients. Again, these are commonly reported in statistical software output. The computed value of  $t$  for temperature on the Excel output is  $-5.934$  and has a  $p$ -value of  $0.000$ . Because the  $p$ -value is less than  $0.05$ , the regression coefficient for the independent variable temperature is not equal to zero and should be included in the equation to predict heating costs. For insulation, the value of  $t$  is  $-3.119$  and has a  $p$ -value of  $0.007$ . As with temperature, the  $p$ -value is less than  $0.05$ , so we conclude that the insulation regression coefficient is not equal to zero and should be included in the equation to predict heating cost. In contrast to temperature and insulation, the  $p$ -value to test the “age of the furnace” regression coefficient is  $0.148$ . It is clearly greater than  $0.05$ , so we conclude that the “age of furnace” regression coefficient could equal 0. Further, as an independent variable, it is not a significant predictor of heating cost. Thus, age of furnace should not be included in the equation to predict heating costs.

At this point, we need to develop a strategy for deleting independent variables. In the Salsberry Realty case, there were three independent variables. For the age of the furnace variable, we failed to reject the null hypothesis that the regression coefficient was zero. It is clear that we should drop that variable and rerun the regression equation. Below is the Excel output where heating cost is the dependent variable and outside temperature and amount of insulation are the independent variables.

|    | A    | B    | C     | D   | E | F  | G       | H          | I         | J      | K              |
|----|------|------|-------|-----|---|--|---------|------------|-----------|--------|----------------|
| 1  | Cost | Temp | Insul | Age |   | SUMMARY OUTPUT                             |         |            |           |        |                |
| 2  | 250  | 35   | 3     | 6   |   |  |         |            |           |        |                |
| 3  | 360  | 29   | 4     | 10  |   | Regression Statistics                      |         |            |           |        |                |
| 4  | 165  | 36   | 7     | 3   |   | Multiple R                                 | 0.881   |            |           |        |                |
| 5  | 43   | 60   | 6     | 9   |   | R Square                                   | 0.776   |            |           |        |                |
| 6  | 92   | 65   | 5     | 6   |   | Adjusted R Square                          | 0.749   |            |           |        |                |
| 7  | 200  | 30   | 5     | 5   |   | Standard Error                             | 52.982  |            |           |        |                |
| 8  | 355  | 10   | 6     | 7   |   | Observations                               | 20      |            |           |        |                |
| 9  | 290  | 7    | 10    | 10  |   |  |         |            |           |        |                |
| 10 | 230  | 21   | 9     | 11  |   | ANOVA                                      |         |            |           |        |                |
| 11 | 120  | 55   | 2     | 5   |   |  | df      | SS         | MS        | F      | Significance F |
| 12 | 73   | 54   | 12    | 4   |   | Regression                                 | 2       | 165194.521 | 82597.261 | 29.424 | 0.000          |
| 13 | 205  | 48   | 5     | 1   |   | Residual                                   | 17      | 47721.229  | 2807.131  |        |                |
| 14 | 400  | 20   | 5     | 15  |   | Total                                      | 19      | 212915.750 |           |        |                |
| 15 | 320  | 39   | 4     | 7   |   |  |         |            |           |        |                |
| 16 | 72   | 60   | 8     | 6   |   | Coefficients Standard Error t Stat P-value |         |            |           |        |                |
| 17 | 272  | 20   | 5     | 8   |   | Intercept                                  | 490.286 | 44.410     | 11.040    | 0.000  |                |
| 18 | 94   | 58   | 7     | 3   |   | Temp                                       | -5.150  | 0.702      | -7.337    | 0.000  |                |
| 19 | 190  | 40   | 8     | 11  |   | Insul                                      | -14.718 | 4.934      | -2.983    | 0.008  |                |
| 20 | 235  | 27   | 9     | 8   |   |  |         |            |           |        |                |

Source: Microsoft Excel

Summarizing the results from this new output:

1. The new regression equation is:

$$\hat{y} = 490.286 - 5.150x_1 - 14.718x_2$$

Notice that the regression coefficients for outside temperature ( $x_1$ ) and amount of insulation ( $x_2$ ) are similar to but not exactly the same as when we included the independent variable age of the furnace. Compare the above equation to that in the Excel output earlier in this section. Both of the regression coefficients are negative as in the earlier equation.

2. The details of the global test are as follows:

$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_1 : \text{Not all of the } \beta_i\text{'s} = 0$$

The  $F$  distribution is the test statistic and there are  $k = 2$  degrees of freedom in the numerator and  $n - (k + 1) = 20 - (2 + 1) = 17$  degrees of freedom in the denominator. Using the .05 significance level and Appendix B.6A, the decision rule is to reject  $H_0$  if  $F$  is greater than 3.59. We compute the value of  $F$  as follows:

$$F = \frac{\text{SSR}/k}{\text{SSE}/[n - (k + 1)]} = \frac{165,194.521/2}{47,721.229/[20 - (2 + 1)]} = 29.424$$

Because the computed value of  $F$  (29.424) is greater than the critical value (3.59), the null hypothesis is rejected and the alternate accepted. We conclude that at least one of the regression coefficients is different from 0.

Using the  $p$ -value, the  $F$  test statistic (29.424) has a  $p$ -value (0.000) which is clearly less than 0.05. Therefore, we reject the null hypothesis and accept the alternate. We conclude that at least one of the regression coefficients is different from 0.

3. The next step is to conduct a test of the regression coefficients individually. We want to determine if one or both of the regression coefficients are different from 0. The null and alternate hypotheses for each of the independent variables are:

|                     |            |
|---------------------|------------|
| Outside Temperature | Insulation |
|---------------------|------------|

$$H_0 : \beta_1 = 0$$

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_1 \neq 0$$

$$H_1 : \beta_2 \neq 0$$

The test statistic is the  $t$  distribution with  $n - (k + 1) = 20 - (2 + 1) = 17$  degrees of freedom. Using the .05 significance level and Appendix B.5, the decision rule is to reject  $H_0$  if the computed value of  $t$  is less than  $-2.110$  or greater than  $2.110$ .

|                     |            |
|---------------------|------------|
| Outside Temperature | Insulation |
|---------------------|------------|

$$t = \frac{b_1 - 0}{s_{b_1}} = \frac{-5.150 - 0}{0.702} = -7.337 \quad t = \frac{b_2 - 0}{s_{b_2}} = \frac{-14.718 - 0}{4.934} = -2.983$$

In both tests, we reject  $H_0$  and accept  $H_1$ . We conclude that each of the regression coefficients is different from 0. Both outside temperature and amount of insulation are useful variables in explaining the variation in heating costs.

Using  $p$ -values, the  $p$ -value for the temperature  $t$ -statistic is 0.000 and the  $p$ -value for the insulation  $t$ -statistic is 0.008. Both  $p$ -values are less than 0.05, so in both tests we reject the null hypothesis and conclude that each of the regression coefficients is different from 0. Both outside temperature and amount of insulation are useful variables in explaining the variation in heating costs.

In the heating cost example, it was clear which independent variable to delete. However, in some instances, which variable to delete may not be as clear-cut. To explain, suppose we develop a multiple regression equation based on five independent variables. We conduct the global test and find that some of the regression coefficients

are different from zero. Next, we test the regression coefficients individually and find that three are significant and two are not. The preferred procedure is to drop the single independent variable with the *smallest absolute t-value* or *largest p-value* and rerun the regression equation with the four remaining variables, then, on the new regression equation with four independent variables, conduct the individual tests. If there are still regression coefficients that are not significant, again drop the variable with the smallest absolute *t-value* or the largest, nonsignificant *p-value*. To describe the process in another way, we should delete only one variable at a time. Each time we delete a variable, we need to rerun the regression equation and check the remaining variables.

This process of selecting variables to include in a regression model can be automated using Excel, Minitab, MegaStat, or other statistical software. Most of the software systems include methods to sequentially remove and/or add independent variables and at the same time provide estimates of the percentage of variation explained (the *R-square* term). Two of the common methods are **stepwise regression** and **best subset regression**. It may take a long time, but in the extreme we could compute every regression between the dependent variable and any possible subset of the independent variables.

Unfortunately, on occasion, the software may work “too hard” to find an equation that fits all the quirks of your particular data set. The suggested equation may not represent the relationship in the population. Judgment is needed to choose among the equations presented. Consider whether the results are logical. They should have a simple interpretation and be consistent with your knowledge of the application under study.

**SELF-REVIEW 14-3**



The regression output about eating places in Myrtle Beach is repeated below (see earlier self-reviews).

| Predictor      | Coefficient | SE Coefficient | t       | p-value |
|----------------|-------------|----------------|---------|---------|
| Constant       | 2.50        | 1.50           | 1.667   | 0.111   |
| x <sub>1</sub> | 3.00        | 1.50           | 2.000   | 0.056   |
| x <sub>2</sub> | 4.00        | 3.00           | 1.333   | 0.194   |
| x <sub>3</sub> | -3.00       | 0.20           | -15.000 | 0.000   |
| x <sub>4</sub> | 0.20        | 0.05           | 4.000   | 0.000   |
| x <sub>5</sub> | 1.00        | 1.50           | 0.667   | 0.511   |

| Analysis of Variance |    |     |    |    |         |
|----------------------|----|-----|----|----|---------|
| Source               | DF | SS  | MS | F  | p-value |
| Regression           | 5  | 100 | 20 | 10 | 0.000   |
| Residual Error       | 20 | 40  | 2  |    |         |
| Total                | 25 | 140 |    |    |         |

- (a) Perform a global test of hypothesis to check if any of the regression coefficients are different from 0. What do you decide? Use the .05 significance level.
- (b) Do an individual test of each independent variable. Which variables would you consider eliminating? Use the .05 significance level.
- (c) Outline a plan for possibly removing independent variables.

**EXERCISES**

7. Given the following regression output,

| Predictor      | Coefficient | SE Coefficient | t     | p-value |
|----------------|-------------|----------------|-------|---------|
| Constant       | 84.998      | 1.863          | 45.62 | 0.000   |
| x <sub>1</sub> | 2.391       | 1.200          | 1.99  | 0.051   |
| x <sub>2</sub> | -0.409      | 0.172          | -2.38 | 0.021   |

| Analysis of Variance |    |         |        |       |         |
|----------------------|----|---------|--------|-------|---------|
| Source               | DF | SS      | MS     | F     | p-value |
| Regression           | 2  | 77.907  | 38.954 | 4.138 | 0.021   |
| Residual Error       | 62 | 583.693 | 9.414  |       |         |
| Total                | 64 | 661.600 |        |       |         |

answer the following questions.

- a. Write the regression equation.
  - b. If  $x_1$  is 4 and  $x_2$  is 11, what is the expected or predicted value of the dependent variable?
  - c. How large is the sample? How many independent variables are there?
  - d. Conduct a global test of hypothesis to see if any of the set of regression coefficients could be different from 0. Use the .05 significance level. What is your conclusion?
  - e. Conduct a test of hypothesis for each independent variable. Use the .05 significance level. Which variable would you consider eliminating?
  - f. Outline a strategy for deleting independent variables in this case.
8. The following regression output was obtained from a study of architectural firms. The dependent variable is the total amount of fees in millions of dollars.

| Predictor | Coefficient | SE Coefficient | t      | p-value |
|-----------|-------------|----------------|--------|---------|
| Constant  | 7.987       | 2.967          | 2.690  | 0.010   |
| $x_1$     | 0.122       | 0.031          | 3.920  | 0.000   |
| $x_2$     | -1.220      | 0.053          | -2.270 | 0.028   |
| $x_3$     | -0.063      | 0.039          | -1.610 | 0.114   |
| $x_4$     | 0.523       | 0.142          | 3.690  | 0.001   |
| $x_5$     | -0.065      | 0.040          | -1.620 | 0.112   |

| Analysis of Variance |    |         |       |       |         |
|----------------------|----|---------|-------|-------|---------|
| Source               | DF | SS      | MS    | F     | p-value |
| Regression           | 5  | 371.000 | 74.2  | 12.89 | 0.000   |
| Residual Error       | 46 | 2647.38 | 57.55 |       |         |
| Total                | 51 | 6357.38 |       |       |         |

$x_1$  is the number of architects employed by the company.

$x_2$  is the number of engineers employed by the company.

$x_3$  is the number of years involved with health care projects.

$x_4$  is the number of states in which the firm operates.

$x_5$  is the percent of the firm's work that is health care–related.

- a. Write out the regression equation.
- b. How large is the sample? How many independent variables are there?
- c. Conduct a global test of hypothesis to see if any of the set of regression coefficients could be different from 0. Use the .05 significance level. What is your conclusion?
- d. Conduct a test of hypothesis for each independent variable. Use the .05 significance level. Which variable would you consider eliminating first?
- e. Outline a strategy for deleting independent variables in this case.

#### LO14-4

Evaluate the assumptions of multiple regression.

## EVALUATING THE ASSUMPTIONS OF MULTIPLE REGRESSION

In the previous section, we described the methods to statistically evaluate the multiple regression equation. The results of the test let us know if at least one of the coefficients was not equal to zero, and we described a procedure of evaluating each regression coefficient. We also discussed the decision-making process for including and excluding independent variables in the multiple regression equation.

It is important to know that the validity of the statistical global and individual tests relies on several assumptions. So if the assumptions are not true, the results might be biased or misleading. However, strict adherence to the following assumptions is not always possible. Fortunately, the statistical techniques discussed in this chapter are robust enough to work effectively even when one or more of the assumptions are violated. Even if the values in the multiple regression equation are “off” slightly, our estimates using a multiple regression equation will be closer than any that could be made otherwise.

In Chapter 13, we listed the necessary assumptions for regression when we considered only a single independent variable. The assumptions for multiple regression are similar.

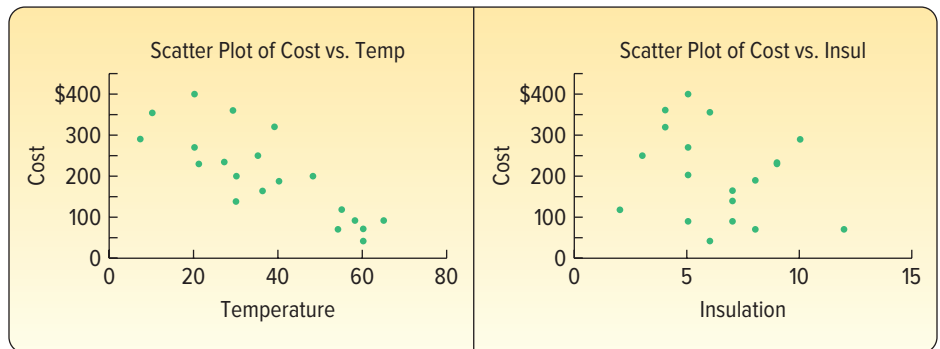
1. **There is a linear relationship.** That is, there is a straight-line relationship between the dependent variable and the set of independent variables.
2. **The variation in the residuals is the same for both large and small values of  $\hat{y}$ .** To put it another way,  $(y - \hat{y})$  is unrelated to whether  $\hat{y}$  is large or small.
3. **The residuals follow the normal probability distribution.** Recall the residual is the difference between the actual value of  $y$  and the estimated value  $\hat{y}$ . So the term  $(y - \hat{y})$  is computed for every observation in the data set. These residuals should approximately follow a normal probability distribution with a mean of 0.
4. **The independent variables should not be correlated.** That is, we would like to select a set of independent variables that are not themselves correlated.
5. **The residuals are independent.** This means that successive observations of the dependent variable are not correlated. This assumption is often violated when time is involved with the sampled observations.

In this section, we present a brief discussion of each of these assumptions. In addition, we provide methods to validate these assumptions and indicate the consequences if these assumptions cannot be met. For those interested in additional discussion, search on the term “Applied Linear Models.”

## Linear Relationship

Let’s begin with the linearity assumption. The idea is that the relationship between the set of independent variables and the dependent variable is linear. If we are considering two independent variables, we can visualize this assumption. The two independent variables and the dependent variable would form a three-dimensional space. The regression equation would then form a plane as shown on page 420. We can evaluate this assumption with scatter diagrams and residual plots.

**Using Scatter Diagrams** The evaluation of a multiple regression equation should always include a scatter diagram that plots the dependent variable against each independent variable. These graphs help us to visualize the relationships and provide some initial information about the direction (positive or negative), linearity, and strength of the relationship. For example, the scatter diagrams for the home heating example follow. The plots suggest a fairly strong negative, linear relationship between heating cost and temperature, and a negative relationship between heating cost and insulation.



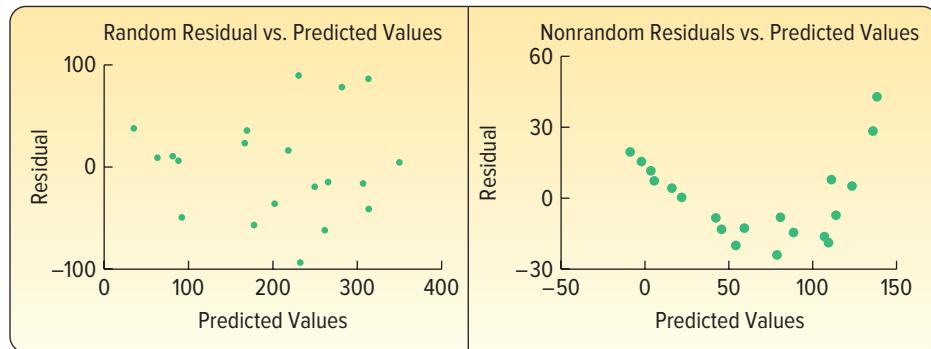
**Using Residual Plots** Recall that a residual  $(y - \hat{y})$  can be computed using the multiple regression equation for each observation in a data set. In Chapter 13, we discussed

the idea that the best regression line passed through the center of the data in a scatter plot. In this case, you would find a good number of the observations above the regression line (these residuals would have a positive sign) and a good number of the observations below the line (these residuals would have a negative sign). Further, the observations would be scattered above and below the line over the entire range of the independent variable.

The same concept is true for multiple regression, but we cannot graphically portray the multiple regression. However, plots of the residuals can help us evaluate the linearity of the multiple regression equation. To investigate, the residuals are plotted on the vertical axis against the predicted variable,  $\hat{y}$ . In the following graphs, the left graph shows the residual plots for the home heating cost example. Notice the following:

- The residuals are plotted on the vertical axis and are centered around zero. There are both positive and negative residuals.
- The residual plots show a random distribution of positive and negative values across the entire range of the variable plotted on the horizontal axis.
- The points are scattered and there is no obvious pattern, so there is no reason to doubt the linearity assumption.

The plot on the right shows nonrandom residuals. See that the residual plot does *not* show a random distribution of positive and negative values across the entire range of the variable plotted on the horizontal axis. In fact, the graph shows a nonlinear pattern of the residuals. This indicates the relationship is probably not linear. In this case, we would evaluate different transformations of the variables in the equation, as discussed in Chapter 13.



## Variation in Residuals Same for Large and Small $\hat{y}$ Values

This requirement indicates that the variation in the residuals is constant, regardless of whether the predicted values are large or small. To cite a specific example which may violate the assumption, suppose we use the single independent variable age to explain variation in monthly income. We suspect that as age increases so does income, but it also seems reasonable that as age increases there may be more variation around the regression line. That is, there will likely be more variation in income for 50-year-olds than for 35-year-olds. The requirement for constant variation around the regression line is called **homoscedasticity**.

**HOMOSCEDASTICITY** The variation around the regression equation is the same for all of the values of the independent variables.

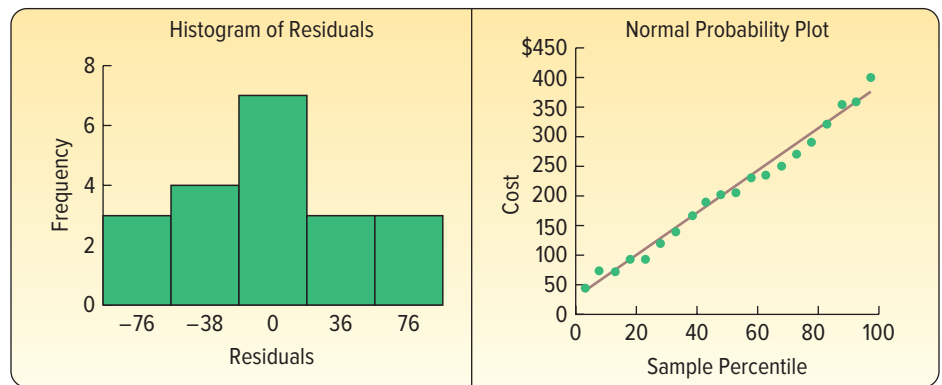
To check for homoscedasticity, the residuals are plotted against  $\hat{y}$ . This is the same graph that we used to evaluate the assumption of linearity. Based on the scatter diagram, it is reasonable to conclude that this assumption has not been violated.

## Distribution of Residuals

To be sure that the inferences we make in the global and individual hypothesis tests are valid, we evaluate the distribution of residuals. Ideally, the residuals should follow a normal probability distribution.

To evaluate this assumption, we can organize the residuals into a frequency distribution. The Histogram of Residuals graph is shown on the left for the home heating cost example. Although it is difficult to show that the residuals follow a normal distribution with only 20 observations, it does appear the normality assumption is reasonable.

Another graph that helps to evaluate the assumption of normally distributed residuals is called a normal probability plot and is shown to the right of the histogram. This graphical analysis is often included in statistical software. If the plotted points are fairly close to a straight line drawn from the lower left to the upper right of the graph, the normal probability plot supports the assumption of normally distributed residuals. This plot supports the assumption of normally distributed residuals.



In this case, both graphs support the assumption that the residuals follow the normal probability distribution. Therefore, the inferences that we made based on the global and individual hypothesis tests are supported with the results of this evaluation.

## Multicollinearity

Multicollinearity exists when independent variables are correlated. Correlated independent variables make it difficult to make inferences about the individual regression coefficients and their individual effects on the dependent variable. In practice, it is nearly impossible to select variables that are completely unrelated. To put it another way, it is nearly impossible to create a set of independent variables that are not correlated to some degree. However, a general understanding of the issue of multicollinearity is important.

First, multicollinearity does not affect a multiple regression equation’s ability to predict the dependent variable. However, when we are interested in evaluating the relationship between each independent variable and the dependent variable, multicollinearity may show unexpected results.

For example, if we use two highly correlated independent variables, high school GPA and high school class rank, to predict the GPA of incoming college freshmen (dependent variable), we would expect that both independent variables would be positively related to the dependent variable. However, because the independent variables are highly correlated, one of the independent variables may have an unexpected and inexplicable negative sign. In essence, these two independent variables are redundant in that they explain the same variation in the dependent variable.

A second reason to avoid correlated independent variables is they may lead to erroneous results in the hypothesis tests for the individual independent variables. This



is due to the instability of the standard error of estimate. Several clues that indicate problems with multicollinearity include the following:

1. An independent variable known to be an important predictor ends up having a regression coefficient that is not significant.
2. A regression coefficient that should have a positive sign turns out to be negative, or vice versa.
3. When an independent variable is added or removed, there is a drastic change in the values of the remaining regression coefficients.

In our evaluation of a multiple regression equation, an approach to reducing the effects of multicollinearity is to carefully select the independent variables that are included in the regression equation. A general rule is if the correlation between two independent variables is between  $-0.70$  and  $0.70$ , there likely is not a problem using both of the independent variables. A more precise test is to use the **variance inflation factor**. It is usually written *VIF*. The value of *VIF* is found as follows:

**VARIANCE INFLATION FACTOR**

$$VIF = \frac{1}{1 - R_j^2} \quad (14-7)$$

The term  $R_j^2$  refers to the coefficient of determination, where the selected *independent variable* is used as a dependent variable and the remaining independent variables are used as independent variables. A *VIF* greater than 10 is considered unsatisfactory, indicating that the independent variable should be removed from the analysis. The following example will explain the details of finding the *VIF*.

**VARIANCE INFLATION FACTOR** A test used to detect correlation among independent variables.

► **EXAMPLE**

Refer to the data in Table 14–1, which relate the heating cost to the independent variables outside temperature, amount of insulation, and age of furnace. Develop a correlation matrix for all the independent variables. Does it appear there is a problem with multicollinearity? Find and interpret the variance inflation factor for each of the independent variables.

**SOLUTION**

We begin by finding the correlation matrix for the dependent variable and the three independent variables. A correlation matrix shows the correlation between all pairs of the variables. A portion of that output follows:

|       | <i>Cost</i> | <i>Temp</i> | <i>Insul</i> | <i>Age</i> |
|-------|-------------|-------------|--------------|------------|
| Cost  | 1.000       |             |              |            |
| Temp  | –0.812      | 1.000       |              |            |
| Insul | –0.257      | –0.103      | 1.000        |            |
| Age   | 0.537       | –0.486      | 0.064        | 1.000      |

The highlighted area indicates the correlation among the independent variables. Because all of the correlations are between  $-.70$  and  $.70$ , we do not suspect problems with multicollinearity. The largest correlation among the independent variables is  $-0.486$ , between age and temperature.

To confirm this conclusion, we compute the *VIF* for each of the three independent variables. We will consider the independent variable temperature first. We use the Regression Analysis in Excel to find the multiple coefficient of determination with temperature as the *dependent variable* and amount of insulation and age of the furnace as independent variables. The relevant regression output follows.

| SUMMARY OUTPUT               |           |           |           |          |                       |
|------------------------------|-----------|-----------|-----------|----------|-----------------------|
| <i>Regression Statistics</i> |           |           |           |          |                       |
| Multiple R                   | 0.491     |           |           |          |                       |
| R Square                     | 0.241     |           |           |          |                       |
| Adjusted R Square            | 0.152     |           |           |          |                       |
| Standard Error               | 16.031    |           |           |          |                       |
| Observations                 | 20        |           |           |          |                       |
| <i>ANOVA</i>                 |           |           |           |          |                       |
|                              | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
| Regression                   | 2         | 1390.291  | 695.145   | 2.705    | 0.096                 |
| Residual                     | 17        | 4368.909  | 256.995   |          |                       |
| Total                        | 19        | 5759.200  |           |          |                       |

The coefficient of determination is .241, so inserting this value into the *VIF* formula:

$$VIF = \frac{1}{1 - R_1^2} = \frac{1}{1 - .241} = 1.32$$

The *VIF* value of 1.32 is less than the upper limit of 10. This indicates that the independent variable temperature is not strongly correlated with the other independent variables.

Again, to find the *VIF* for insulation, we would develop a regression equation with insulation as the *dependent variable* and temperature and age of furnace as independent variables. For this equation, the  $R^2$  is .011 and, using formula (14–7), the *VIF* for insulation would be 1.011. To find the *VIF* for age, we would develop a regression equation with age as the dependent variable and temperature and insulation as the independent variables. For this equation, the  $R^2$  is .236 and, using formula (14–7), the *VIF* for age would be 1.310. All the *VIF* values are less than 10. Hence, we conclude there is not a problem with multicollinearity in this example.

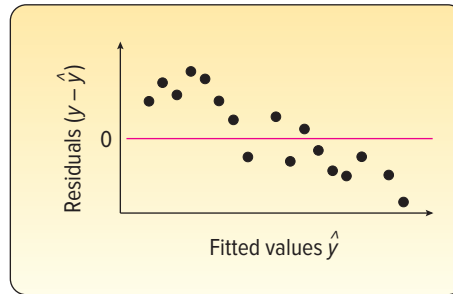
## Independent Observations

The fifth assumption about regression and correlation analysis is that successive residuals should be independent. This means that there is not a pattern to the residuals, the residuals are not highly correlated, and there are not long runs of positive or negative residuals. When successive residuals are correlated, we refer to this condition as **autocorrelation**.

**AUTOCORRELATION** Successive residuals in a time series are correlated.

Autocorrelation frequently occurs when the data are collected over a period of time. For example, we wish to predict yearly sales of Agis Software Inc. based on the time and the amount spent on advertising. The dependent variable is yearly sales and the independent variables are time and amount spent on advertising. It is likely that for a period of time the actual points will be above the regression plane (remember there are two independent variables) and then for a period of time the points will be below the regression plane. The graph below shows the residuals plotted on the vertical axis and

the fitted values  $\hat{y}$  on the horizontal axis. Note the run of residuals above the mean of the residuals, followed by a run below the mean. A scatter plot such as this would indicate possible autocorrelation.



**LO14-5**

Use and interpret a qualitative, dummy variable in multiple regression.

**STATISTICS IN ACTION**

Multiple regression has been used in a variety of legal proceedings. It is particularly useful in cases alleging discrimination by gender or race. As an example, suppose that a woman alleges that Company X's wage rates are unfair to women. To support the claim, the plaintiff produces data showing that, on the average, women earn less than men. In response, Company X argues that its wage rates are based on experience, training, and skill and that its female employees, on the average, are younger and less experienced than the male employees. In fact, the company might further argue that the current situation is actually due to its recent successful efforts to hire more women.

## QUALITATIVE INDEPENDENT VARIABLES

In the previous example/solution regarding heating cost, the two independent variables outside temperature and insulation were quantitative; that is, they were numerical in nature. Frequently, we wish to use nominal-scale variables—if a home has a swimming pool, or whether the sports team was the home or the visiting team—in our analysis. These are called **qualitative variables** because they describe a particular quality or attribute. To use a qualitative variable in regression analysis, we use a scheme of **dummy variables** in which one of the two possible conditions is coded 0 and the other 1.

**DUMMY VARIABLE** A variable in which there are only two possible outcomes. For analysis, one of the outcomes is coded a 1 and the other a 0.

For example, we are interested in estimating an executive's salary on the basis of years of job experience and whether he or she graduated from college. "Graduation from college" can take on only one of two conditions: yes or no. Thus, it is considered a qualitative variable.

Suppose in the Salsberry Realty example that the independent variable "garage" is added. For those homes without an attached garage, 0 is used; for homes with an attached garage, a 1 is used. We will refer to the "garage" variable as  $x_4$ . The data from Table 14–2 are entered into the Excel system. Recall that the variable "age of the furnace" is not included in the analysis because we determined that it was not significantly related to heating cost.

The output from Excel is:

|    | A    | B    | C     | D      | E | F   | G         | H          | I         | J        | K                     |
|----|------|------|-------|--------|---|---|-----------|------------|-----------|----------|-----------------------|
| 1  | Cost | Temp | Insul | Garage |   | SUMMARY OUTPUT                                    |           |            |           |          |                       |
| 2  | 250  | 35   | 3     | 0      |   |   |           |            |           |          |                       |
| 3  | 360  | 29   | 4     | 1      |   | <i>Regression Statistics</i>                      |           |            |           |          |                       |
| 4  | 165  | 36   | 7     | 0      |   | Multiple R  | 0.933     |            |           |          |                       |
| 5  | 43   | 60   | 6     | 0      |   | R Square  | 0.870     |            |           |          |                       |
| 6  | 92   | 65   | 5     | 0      |   | Adjusted R Square                                 | 0.845     |            |           |          |                       |
| 7  | 200  | 30   | 5     | 0      |   | Standard Error                                    | 41.618    |            |           |          |                       |
| 8  | 355  | 10   | 6     | 1      |   | Observations                                      | 20        |            |           |          |                       |
| 9  | 290  | 7    | 10    | 1      |   | <i>ANOVA</i>                                      |           |            |           |          |                       |
| 10 | 230  | 21   | 9     | 0      |   |   | <i>df</i> | <i>SS</i>  | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
| 11 | 120  | 55   | 2     | 0      |   | Regression  | 3         | 185202.269 | 61734.090 | 35.641   | 0.000                 |
| 12 | 73   | 54   | 12    | 0      |   | Residual  | 16        | 27713.481  | 1732.093  |          |                       |
| 13 | 205  | 48   | 5     | 1      |   | Total   | 19        | 212915.750 |           |          |                       |
| 14 | 400  | 20   | 5     | 1      |   |   |           |            |           |          |                       |
| 15 | 320  | 39   | 4     | 1      |   |   |           |            |           |          |                       |
| 16 | 72   | 60   | 8     | 0      |   | <i>Coefficients Standard Error t Stat P-value</i> |           |            |           |          |                       |
| 17 | 272  | 20   | 5     | 1      |   | Intercept   | 393.666   | 45.001     | 8.748     | 0.000    |                       |
| 18 | 94   | 58   | 7     | 0      |   | Temp  | -3.963    | 0.653      | -6.072    | 0.000    |                       |
| 19 | 190  | 40   | 8     | 1      |   | Insul   | -11.334   | 4.002      | -2.832    | 0.012    |                       |
| 20 | 235  | 27   | 9     | 0      |   | Garage  | 77.432    | 22.783     | 3.399     | 0.004    |                       |

Source: Microsoft Excel

**TABLE 14-2** Home Heating Costs, Temperature, Insulation, and Presence of a Garage for a Sample of 20 Homes

| Cost,<br>$y$ | Temperature,<br>$x_1$ | Insulation,<br>$x_2$ | Garage,<br>$x_4$ |
|--------------|-----------------------|----------------------|------------------|
| \$250        | 35                    | 3                    | 0                |
| 360          | 29                    | 4                    | 1                |
| 165          | 36                    | 7                    | 0                |
| 43           | 60                    | 6                    | 0                |
| 92           | 65                    | 5                    | 0                |
| 200          | 30                    | 5                    | 0                |
| 355          | 10                    | 6                    | 1                |
| 290          | 7                     | 10                   | 1                |
| 230          | 21                    | 9                    | 0                |
| 120          | 55                    | 2                    | 0                |
| 73           | 54                    | 12                   | 0                |
| 205          | 48                    | 5                    | 1                |
| 400          | 20                    | 5                    | 1                |
| 320          | 39                    | 4                    | 1                |
| 72           | 60                    | 8                    | 0                |
| 272          | 20                    | 5                    | 1                |
| 94           | 58                    | 7                    | 0                |
| 190          | 40                    | 8                    | 1                |
| 235          | 27                    | 9                    | 0                |
| 139          | 30                    | 7                    | 0                |

What is the effect of the garage variable? Should it be included in the analysis? To show the effect of the variable, suppose we have two homes exactly alike next to each other in Buffalo, New York; one has an attached garage and the other does not. Both homes have 3 inches of insulation, and the mean January temperature in Buffalo is 20 degrees. For the house without an attached garage, a 0 is substituted for  $x_4$  in the regression equation. The estimated heating cost is \$280.404, found by:

$$\begin{aligned} \hat{y} &= 393.666 - 3.963x_1 - 11.334x_2 + 77.432x_4 \\ &= 393.666 - 3.963(20) - 11.334(3) + 77.432(0) = 280.404 \end{aligned}$$

For the house with an attached garage, a 1 is substituted for  $x_4$  in the regression equation. The estimated heating cost is \$357.836, found by:

$$\begin{aligned} \hat{y} &= 393.666 - 3.963x_1 - 11.334x_2 + 77.432x_4 \\ &= 393.666 - 3.963(20) - 11.334(3) + 77.432(1) = 357.836 \end{aligned}$$

The difference between the estimated heating costs is \$77.432 (\$357.836 – \$280.404). Hence, we can expect the cost to heat a house with an attached garage to be \$77.432 more than the cost for an equivalent house without a garage.

We have shown the difference between the two types of homes to be \$77.432, but is the difference significant? We conduct the following test of hypothesis.

$$\begin{aligned} H_0: \beta_4 &= 0 \\ H_1: \beta_4 &\neq 0 \end{aligned}$$

The information necessary to answer this question is in the output at the bottom of the previous page. The regression coefficient for the independent variable garage is \$77.432, and the standard deviation of the sampling distribution is 22.783. We identify this as the fourth independent variable, so we use a subscript of 4. (Remember we

dropped age of the furnace, the third independent variable.) Finally, we insert these values in formula (14–6).

$$t = \frac{b_4 - 0}{s_{b_4}} = \frac{77.432 - 0}{22.783} = 3.399$$

There are three independent variables in the analysis, so there are  $n - (k + 1) = 20 - (3 + 1) = 16$  degrees of freedom. The critical value from Appendix B.5 is 2.120. The decision rule, using a two-tailed test and the .05 significance level, is to reject  $H_0$  if the computed  $t$  is to the left of  $-2.120$  or to the right of 2.120. Because the computed value of 3.399 is to the right of 2.120, the null hypothesis is rejected. We conclude that the regression coefficient is not zero. The independent variable garage should be included in the analysis.

Using the  $p$ -value approach, the computed  $t$ -value of 3.399 has a  $p$ -value of 0.004. This value is less than the .05 significance level. Therefore, we reject the null hypothesis. We conclude that the regression coefficient is not zero and the independent variable garage should be included in the analysis.

Is it possible to use a qualitative variable with more than two possible outcomes? Yes, but the coding scheme becomes more complex and will require a series of dummy variables. To explain, suppose a company is studying its sales as they relate to advertising expense by quarter for the last 5 years. Let sales be the dependent variable and advertising expense be the first independent variable,  $x_1$ . To include the qualitative information regarding the quarter, we use three additional independent variables. For the variable  $x_2$ , the five observations referring to the first quarter of each of the 5 years are coded 1 and the other quarters 0. Similarly, for  $x_3$  the five observations referring to the second quarter are coded 1 and the other quarters 0. For  $x_4$ , the five observations referring to the third quarter are coded 1 and the other quarters 0. An observation that does not refer to any of the first three quarters must refer to the fourth quarter, so a distinct independent variable referring to this quarter is not necessary.

### SELF-REVIEW 14-4



A study by the American Realtors Association investigated the relationship between the commissions earned by sales associates last year and the number of months since the associates earned their real estate licenses. Also of interest in the study is the gender of the sales associate. Below is a portion of the regression output. The dependent variable is commissions, which is reported in \$000, and the independent variables are months since the license was earned and gender (female = 1 and male = 0).

*Regression Analysis*

*Regression Statistics*

|                   |       |
|-------------------|-------|
| Multiple R        | 0.801 |
| R Square          | 0.642 |
| Adjusted R Square | 0.600 |
| Standard Error    | 3.219 |
| Observations      | 20    |

| ANOVA      | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>p-value</i> |
|------------|-----------|-----------|-----------|----------|----------------|
| Regression | 2         | 315.9291  | 157.9645  | 15.2468  | 0.0002         |
| Residual   | 17        | 176.1284  | 10.36049  |          |                |
| Total      | 19        | 492.0575  |           |          |                |

|           | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>p-value</i> |
|-----------|---------------------|-----------------------|---------------|----------------|
| Intercept | 15.7625             | 3.0782                | 5.121         | .0001          |
| Months    | 0.4415              | 0.0839                | 5.262         | .0001          |
| Gender    | 3.8598              | 1.4724                | 2.621         | .0179          |

- (a) Write out the regression equation. How much commission would you expect a female agent to make who earned her license 30 months ago?
- (b) Do the female agents on average make more or less than the male agents? How much more?
- (c) Conduct a test of hypothesis to determine if the independent variable gender should be included in the analysis. Use the .05 significance level. What is your conclusion?

**LO14-6**

Apply stepwise regression to develop a multiple regression model.

## STEPWISE REGRESSION

In our heating cost example (see sample information in Table 14–1), we considered three independent variables: the mean outside temperature, the amount of insulation in the home, and the age of the furnace. To obtain the equation, we first ran a global or “all at once” test to determine if any of the regression coefficients were significant. When we found at least one to be significant, we tested the regression coefficients individually to determine which were important. We kept the independent variables that had significant regression coefficients and left the others out. By retaining the independent variables with significant coefficients, we found the regression equation that used the fewest independent variables. This made the regression equation easier to interpret. Then we considered the qualitative variable garage and found that it was significantly related to heating cost. The variable garage was added to the equation.

Deciding which set of independent variables to include in a multiple regression equation can be accomplished using a technique called **stepwise regression**. This technique efficiently builds an equation that only includes independent variables with significant regression coefficients.

**STEPWISE REGRESSION** A step-by-step method to determine a regression equation that begins with a single independent variable and adds or deletes independent variables one by one. Only independent variables with nonzero regression coefficients are included in the regression equation.

In the stepwise method, we develop a sequence of equations. The first equation contains only one independent variable. However, this independent variable is the one from the set of proposed independent variables that explains the most variation in the dependent variable. Stated differently, if we compute all the simple correlations between each independent variable and the dependent variable, the stepwise method first selects the independent variable with the strongest correlation with the dependent variable.

Next, the stepwise method looks at the remaining independent variables and selects the one that will explain the largest percentage of the variation yet unexplained. We continue this process until all the independent variables with significant regression coefficients are included in the regression equation. The advantages to the stepwise method are:

1. Only independent variables with significant regression coefficients are entered into the equation.
2. The steps involved in building the regression equation are clear.
3. It is efficient in finding the regression equation with only significant regression coefficients.
4. The changes in the multiple standard error of estimate and the coefficient of determination are shown.

Stepwise regression procedures are included in many statistical software packages. For example, Minitab’s stepwise regression analysis for the home heating cost problem follows. Note that the final equation, which is reported in the column labeled 3, includes the independent variables temperature, garage, and insulation. These are the same independent variables that were included in our equation using the global test and the test for individual independent variables. The independent variable age, indicating the furnace’s age, is not included because it is not a significant predictor of cost.

| ↓  | C1   | C2   | C3    | C4     |
|----|------|------|-------|--------|
|    | Cost | Temp | Insul | Garage |
| 1  | 250  | 35   | 3     | 0      |
| 2  | 360  | 29   | 4     | 1      |
| 3  | 165  | 36   | 7     | 0      |
| 4  | 43   | 60   | 6     | 0      |
| 5  | 92   | 65   | 5     | 0      |
| 6  | 200  | 30   | 5     | 0      |
| 7  | 355  | 10   | 6     | 1      |
| 8  | 290  | 7    | 10    | 1      |
| 9  | 230  | 21   | 9     | 0      |
| 10 | 120  | 55   | 2     | 0      |
| 11 | 73   | 54   | 12    | 0      |
| 12 | 205  | 48   | 5     | 1      |
| 13 | 400  | 20   | 5     | 1      |
| 14 | 320  | 39   | 4     | 1      |
| 15 | 72   | 60   | 8     | 0      |
| 16 | 272  | 20   | 5     | 1      |
| 17 | 94   | 58   | 7     | 0      |
| 18 | 190  | 40   | 8     | 1      |
| 19 | 235  | 27   | 9     | 0      |
| 20 | 139  | 30   | 7     | 0      |

| Stepwise Regression: Cost versus Temp, Insul, Garage |       |       |       |  |
|--|-------|-------|-------|--|
| Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15           |       |       |       |  |
| Response is Cost on 3 predictors, with N = 20        |       |       |       |  |
| Step   | 1     | 2     | 3     |  |
| Constant   | 388.8 | 300.3 | 393.7 |  |
| Temp   | -4.93 | -3.56 | -3.96 |  |
| T-Value  | -5.89 | -4.70 | -6.07 |  |
| P-Value  | 0.000 | 0.000 | 0.000 |  |
| Garage   |       | 93    | 77    |  |
| T-Value  |       | 3.56  | 3.40  |  |
| P-Value  |       | 0.002 | 0.004 |  |
| Insul  |       |       | -11.3 |  |
| T-Value  |       |       | -2.83 |  |
| P-Value  |       |       | 0.012 |  |
| S  | 63.6  | 49.5  | 41.6  |  |
| R-Sq   | 65.85 | 80.46 | 86.98 |  |
| R-Sq (adj)   | 63.96 | 78.16 | 84.54 |  |
| Mallows Cp   | 26.0  | 10.0  | 4.0   |  |

Source: Minitab

Reviewing the steps and interpreting output:

1. The stepwise procedure selects the independent variable temperature first. This variable explains more of the variation in heating cost than any of the other three proposed independent variables. Temperature explains 65.85% of the variation in heating cost. The regression equation is:

$$\hat{y} = 388.8 - 4.93x_1$$

There is an inverse relationship between heating cost and temperature. For each degree the temperature increases, heating cost is reduced by \$4.93.

2. The next independent variable to enter the regression equation is garage. When this variable is added to the regression equation, the coefficient of determination is increased from 65.85% to 80.46%. That is, by adding garage as an independent variable, we increase the coefficient of determination by 14.61 percentage points. The regression equation after step 2 is:

$$\hat{y} = 300.3 - 3.56x_1 + 93.0x_2$$

Usually the regression coefficients will change from one step to the next. In this case, the coefficient for temperature retained its negative sign, but it changed from  $-4.93$  to  $-3.56$ . This change is reflective of the added influence of the independent variable garage. Why did the stepwise method select the independent variable garage instead of either insulation or age? The increase in  $R^2$ , the coefficient of determination, is larger if garage is included rather than either of the other two variables.

3. At this point, there are two unused variables remaining, insulation and age. Notice on the third step the procedure selects insulation and then stops. This indicates the variable insulation explains more of the remaining variation in heating cost than the age variable does. After the third step, the regression equation is:

$$\hat{y} = 393.7 - 3.96x_1 + 77.0x_2 - 11.3x_3$$

At this point, 86.98% of the variation in heating cost is explained by the three independent variables temperature, garage, and insulation. This is the same  $R^2$  value and regression equation we found on page 442 except for rounding differences.

4. Here, the stepwise procedure stops. This means the independent variable age does not add significantly to the coefficient of determination.

The stepwise method developed the same regression equation, selected the same independent variables, and found the same coefficient of determination as the global and individual tests described earlier in the chapter. The advantage to the stepwise method is that it is more direct than using a combination of the global and individual procedures.

Other methods of variable selection are available. The stepwise method is also called the **forward selection method** because we begin with no independent variables and add one independent variable to the regression equation at each iteration. There is also the **backward elimination method**, which begins with the entire set of variables and eliminates one independent variable at each iteration.

The methods described so far look at one variable at a time and decide whether to include or eliminate that variable. Another approach is the **best-subset regression**. With this method, we look at the best model using one independent variable, the best model using two independent variables, the best model with three, and so on. The criterion is to find the model with the largest  $R^2$  value, regardless of the number of independent variables. Also, each independent variable does not necessarily have a nonzero regression coefficient. Since each independent variable could either be included or not included, there are  $2^k - 1$  possible models, where  $k$  refers to the number of independent variables. In our heating cost example, we considered four independent variables so there are 15 possible regression models, found by  $2^4 - 1 = 16 - 1 = 15$ . We would examine all regression models using one independent variable, all combinations using two variables, all combinations using three independent variables, and the possibility of using all four independent variables. The advantage to the best-subset method is it may examine combinations of independent variables not considered in the stepwise method. The process is available in Minitab and MegaStat.

## EXERCISES

9. **FILE** The manager of High Point Sofa and Chair, a large furniture manufacturer located in North Carolina, is studying the job performance ratings of a sample of 15 electrical repairmen employed by the company. An aptitude test is required by the human resources department to become an electrical repairman. The manager was able to get the score for each repairman in the sample. In addition, he determined which of the repairmen were union members (code = 1) and which were not (code = 0). The sample information is reported below.

| Worker      | Job Performance |                     |                  |
|-------------|-----------------|---------------------|------------------|
|             | Score           | Aptitude Test Score | Union Membership |
| Abbott      | 58              | 5                   | 0                |
| Anderson    | 53              | 4                   | 0                |
| Bender      | 33              | 10                  | 0                |
| Bush        | 97              | 10                  | 0                |
| Center      | 36              | 2                   | 0                |
| Coombs      | 83              | 7                   | 0                |
| Eckstine    | 67              | 6                   | 0                |
| Gloss       | 84              | 9                   | 0                |
| Herd        | 98              | 9                   | 1                |
| Householder | 45              | 2                   | 1                |
| Lori        | 97              | 8                   | 1                |
| Lindstrom   | 90              | 6                   | 1                |
| Mason       | 96              | 7                   | 1                |
| Pierse      | 66              | 3                   | 1                |
| Rohde       | 82              | 6                   | 1                |

- Use a statistical software package to develop a multiple regression equation using the job performance score as the dependent variable and aptitude test score and union membership as independent variables.
- Comment on the regression equation. Be sure to include the coefficient of determination and the effect of union membership. Are these two variables effective in explaining the variation in job performance?
- Conduct a test of hypothesis to determine if union membership should be included as an independent variable.



10. **FILE** A real estate developer wishes to study the relationship between the size of home a client will purchase (in square feet) and other variables. Possible independent variables include the family income, family size, whether there is a senior adult parent living with the family (1 for yes, 0 for no), and the total years of education beyond high school for the husband and wife. The sample information is reported below.

| Family | Square Feet | Income (000s) | Family Size | Senior Parent | Education |
|--------|-------------|---------------|-------------|---------------|-----------|
| 1      | 2,240       | 60.8          | 2           | 0             | 4         |
| 2      | 2,380       | 68.4          | 2           | 1             | 6         |
| 3      | 3,640       | 104.5         | 3           | 0             | 7         |
| 4      | 3,360       | 89.3          | 4           | 1             | 0         |
| 5      | 3,080       | 72.2          | 4           | 0             | 2         |
| 6      | 2,940       | 114           | 3           | 1             | 10        |
| 7      | 4,480       | 125.4         | 6           | 0             | 6         |
| 8      | 2,520       | 83.6          | 3           | 0             | 8         |
| 9      | 4,200       | 133           | 5           | 0             | 2         |
| 10     | 2,800       | 95            | 3           | 0             | 6         |

Develop an appropriate multiple regression equation. Which independent variables would you include in the final regression equation? Use the stepwise method.

#### LO14-7

Apply multiple regression techniques to develop a linear model.

## REVIEW OF MULTIPLE REGRESSION

We described many topics involving multiple regression in this chapter. In this section of the chapter, we focus on a single example with a solution that reviews the procedure and guides your application of multiple regression analysis.

### EXAMPLE

The Bank of New England is a large financial institution serving the New England states as well as New York and New Jersey. The mortgage department of the Bank of New England is studying data from recent loans. Of particular interest is how such factors as the value of the home being purchased (\$000), education level of the head of the household (number of years, beginning with first grade), age of the head of the household, current monthly mortgage payment (in dollars), and gender of the head of the household (male = 1, female = 0) relate to the family income. The mortgage department would like to know whether these variables are effective predictors of family income.

### SOLUTION

**FILE** Consider a random sample of 25 loan applications submitted to the Bank of New England last month. A portion of the sample information is shown in Table 14–3. The entire data set is available in Connect and is identified as Bank of New England.

**TABLE 14–3** Information on Sample of 25 Loans by the Bank of New England

| Loan | Income (\$000) | Value (\$000) | Education | Age | Mortgage | Gender |
|------|----------------|---------------|-----------|-----|----------|--------|
| 1    | 100.7          | 190           | 14        | 53  | 230      | 1      |
| 2    | 99.0           | 121           | 15        | 49  | 370      | 1      |
| 3    | 102.0          | 161           | 14        | 44  | 397      | 1      |
| ⋮    | ⋮              | ⋮             | ⋮         | ⋮   | ⋮        | ⋮      |
| 23   | 102.3          | 163           | 14        | 46  | 142      | 1      |
| 24   | 100.2          | 150           | 15        | 50  | 343      | 0      |
| 25   | 96.3           | 139           | 14        | 45  | 373      | 0      |

We begin by calculating the correlation matrix shown below. It shows the relationship between each of the independent variables and the dependent variable. It helps to identify the independent variables that are more closely related to the dependent variable (family income). The correlation matrix also reveals the independent variables that are highly correlated and possibly redundant.

|           | Income | Value  | Education | Age    | Mortgage | Gender |
|-----------|--------|--------|-----------|--------|----------|--------|
| Income    | 1      |        |           |        |          |        |
| Value     | 0.720  | 1      |           |        |          |        |
| Education | 0.188  | -0.144 | 1         |        |          |        |
| Age       | 0.243  | 0.220  | 0.621     | 1      |          |        |
| Mortgage  | 0.116  | 0.358  | -0.210    | -0.038 | 1        |        |
| Gender    | 0.486  | 0.184  | 0.062     | 0.156  | -0.129   | 1      |

What can we learn from this correlation matrix?

1. The first column shows the correlations between each of the independent variables and the dependent variable family income. Observe that each of the independent variables is positively correlated with family income. The value of the home has the strongest correlation with family income. The level of education of the person applying for the loan and the current mortgage payment have a weak correlation with family income. These two variables are candidates to be dropped from the regression equation.
2. All possible correlations among the independent variables are identified with the green background. Our standard is to look for correlations that exceed an absolute value of .700. None of the independent variables are strongly correlated with each other. This indicates that multicollinearity is not likely.

Next, we compute the multiple regression equation using all the independent variables. The software output follows.

|    | A                            | B                   | C                     | D             | E              | F              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|----------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |                |
| 2  |                              |                     |                       |               |                |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |                |
| 4  | Multiple R                   | 0.866               |                       |               |                |                |
| 5  | R Square                     | 0.750               |                       |               |                |                |
| 6  | Adjusted R Square            | 0.684               |                       |               |                |                |
| 7  | Standard Error               | 1.478               |                       |               |                |                |
| 8  | Observations                 | 25                  |                       |               |                |                |
| 9  |                              |                     |                       |               |                |                |
| 10 | ANOVA                        |                     |                       |               |                |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       | <i>P-value</i> |
| 12 | Regression                   | 5                   | 124.322               | 24.864        | 11.385         | 0.000          |
| 13 | Residual                     | 19                  | 41.494                | 2.184         |                |                |
| 14 | Total                        | 24                  | 165.815               |               |                |                |
| 15 |                              |                     |                       |               |                |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |                |
| 17 | Intercept                    | 70.606              | 7.464                 | 9.459         | 0.000          |                |
| 18 | Value (\$000)                | 0.072               | 0.012                 | 5.769         | 0.000          |                |
| 19 | Education                    | 1.624               | 0.603                 | 2.693         | 0.014          |                |
| 20 | Age                          | -0.122              | 0.078                 | -1.566        | 0.134          |                |
| 21 | Mortgage                     | -0.001              | 0.003                 | -0.319        | 0.753          |                |
| 22 | Gender                       | 1.807               | 0.623                 | 2.901         | 0.009          |                |

Source: Microsoft Excel

The coefficients of determination, that is, both  $R^2$  and adjusted  $R^2$ , are reported at the top of the summary output and highlighted in yellow. The  $R^2$  value is 75.0%, so the five independent variables account for three-quarters of the variation in family income. The adjusted  $R^2$  measures the strength of the relationship between the set of independent variables and family income and also accounts for the number of variables in the regression equation. The adjusted  $R^2$  indicates that the five variables account for 68.4% of the variance of family income. Both of these suggest that the proposed independent variables are useful in predicting family income.

The output also includes the regression equation.

$$\hat{y} = 70.606 + 0.072(\text{Value}) + 1.624(\text{Education}) - 0.122(\text{Age}) - 0.001(\text{Mortgage}) + 1.807(\text{Gender})$$

Be careful in this interpretation. Both income and the value of the home are in thousands of dollars. Here is a summary:

1. An increase of \$1,000 in the value of the home suggests an increase of \$72 in family income. An increase of 1 year of education increases income by \$1,624, another year older reduces income by \$122, and an increase of \$1,000 in the mortgage reduces income by \$1.
2. If a male is head of the household, the value of family income will increase by \$1,807. Remember that “female” was coded 0 and “male” was coded 1, so a male head of household is positively related to home value.
3. The age of the head of household and monthly mortgage payment are inversely related to family income. This is true because the sign of the regression coefficient is negative.

Next we conduct the global hypothesis test. Here we check to see if any of the regression coefficients are different from 0. We use the .05 significance level.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

$$H_1: \text{Not all the } \beta\text{'s are 0}$$

The  $p$ -value from the table (cell F12) is 0.000. Because the  $p$ -value is less than the significance level, we reject the null hypothesis and conclude that at least one of the regression coefficients is not equal to zero.

Next we evaluate the individual regression coefficients. The  $p$ -values to test each regression coefficient are reported in cells E18 through E22 in the software output on the previous page. The null hypothesis and the alternate hypothesis are:

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

The subscript  $i$  represents any particular independent variable. Again using .05 significance levels, the  $p$ -values for the regression coefficients for home value, years of education, and gender are all less than .05. We conclude that these regression coefficients are not equal to zero and are significant predictors of family income. The  $p$ -values for age and mortgage amount are greater than the significance level of .05, so we do not reject the null hypotheses for these variables. The regression coefficients are not different from zero and are not related to family income.

Based on the results of testing each of the regression coefficients, we conclude that the variables age and mortgage amount are not effective predictors of family income. Thus, they should be removed from the multiple regression equation. Remember that we must remove one independent variable at a time and redo the analysis to evaluate the overall effect of removing the variable. Our strategy is to remove the variable with the smallest  $t$ -statistic or the largest  $p$ -value. This variable is mortgage amount. The result of the regression analysis without the mortgage variable follows.

|    | A                            | B                   | C                     | D             | E              | F              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|----------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |                |
| 2  |                              |                     |                       |               |                |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |                |
| 4  | Multiple R                   | 0.865               |                       |               |                |                |
| 5  | R Square                     | 0.748               |                       |               |                |                |
| 6  | Adjusted R Square            | 0.698               |                       |               |                |                |
| 7  | Standard Error               | 1.444               |                       |               |                |                |
| 8  | Observations                 | 25                  |                       |               |                |                |
| 9  |                              |                     |                       |               |                |                |
| 10 | ANOVA                        |                     |                       |               |                |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       | <i>P-value</i> |
| 12 | Regression                   | 4                   | 124.099               | 31.025        | 14.874         | 0.000          |
| 13 | Residual                     | 20                  | 41.716                | 2.086         |                |                |
| 14 | Total                        | 24                  | 165.815               |               |                |                |
| 15 |                              |                     |                       |               |                |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |                |
| 17 | Intercept                    | 70.159              | 7.165                 | 9.791         | 0.000          |                |
| 18 | Value (\$000)                | 0.070               | 0.011                 | 6.173         | 0.000          |                |
| 19 | Education                    | 1.647               | 0.585                 | 2.813         | 0.011          |                |
| 20 | Age                          | -0.122              | 0.076                 | -1.602        | 0.125          |                |
| 21 | Gender                       | 1.846               | 0.596                 | 3.096         | 0.006          |                |

Source: Microsoft Excel

Observe that the  $R^2$  and adjusted  $R^2$  change very little without the mortgage variable. Also observe that the  $p$ -value associated with age is greater than the .05 significance level. So next we remove the age variable and redo the analysis. The regression output with the variables age and mortgage amount removed follows.

|    | A                            | B                   | C                     | D             | E              | F              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|----------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |                |
| 2  |                              |                     |                       |               |                |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |                |
| 4  | Multiple R                   | 0.846               |                       |               |                |                |
| 5  | R Square                     | 0.716               |                       |               |                |                |
| 6  | Adjusted R Square            | 0.676               |                       |               |                |                |
| 7  | Standard Error               | 1.497               |                       |               |                |                |
| 8  | Observations                 | 25                  |                       |               |                |                |
| 9  |                              |                     |                       |               |                |                |
| 10 | ANOVA                        |                     |                       |               |                |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       | <i>P-value</i> |
| 12 | Regression                   | 3                   | 118.743               | 39.581        | 17.658         | 0.000          |
| 13 | Residual                     | 21                  | 47.072                | 2.242         |                |                |
| 14 | Total                        | 24                  | 165.815               |               |                |                |
| 15 |                              |                     |                       |               |                |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |                |
| 17 | Intercept                    | 74.527              | 6.870                 | 10.849        | 0.000          |                |
| 18 | Value (\$000)                | 0.063               | 0.011                 | 5.803         | 0.000          |                |
| 19 | Education                    | 1.016               | 0.449                 | 2.262         | 0.034          |                |
| 20 | Gender                       | 1.770               | 0.616                 | 2.872         | 0.009          |                |

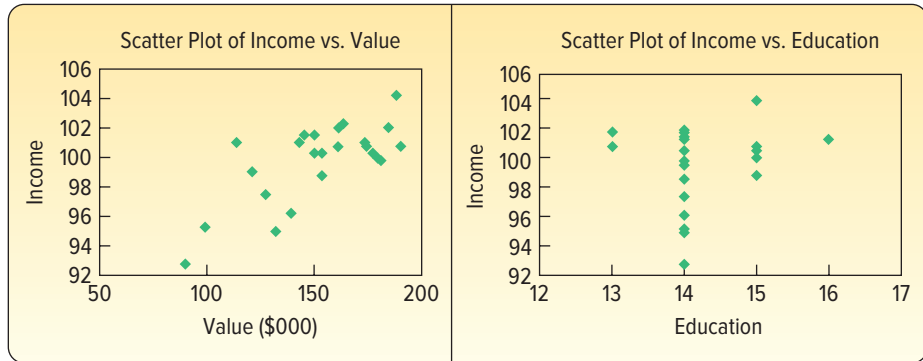
Source: Microsoft Excel

From this output, we conclude:

1. The  $R^2$  and adjusted  $R^2$  values have declined but only slightly. Using all five independent variables, the  $R^2$  value was .750. With the two nonsignificant variables removed, the  $R^2$  and adjusted  $R^2$  values are .716 and .676, respectively. We prefer the equation with the fewer number of independent variables. It is easier to interpret.

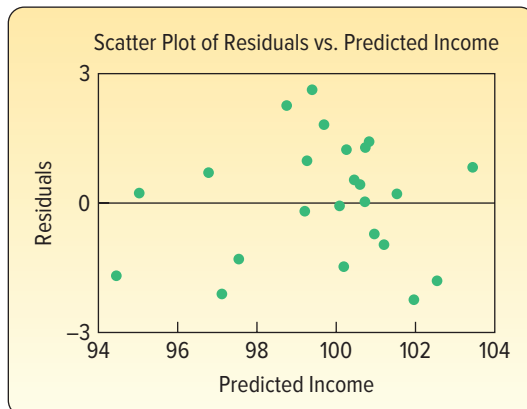
2. In ANOVA, we observe that the  $p$ -value is less than .05. Hence, at least one of the regression coefficients is not equal to zero.
3. Reviewing the significance of the individual coefficients, the  $p$ -values associated with each of the remaining independent variables are less than .05. We conclude that all the regression coefficients are different from zero. Each independent variable is a useful predictor of family income.

Our final step is to examine the regression assumptions (Evaluating the Assumptions of Multiple Regression section on page 436) with our regression model. The first assumption is that there is a linear relationship between each independent variable and the dependent variable. It is not necessary to review the dummy variable gender because there are only two possible outcomes. Below are the scatter plots of family income versus home value and family income versus years of education.



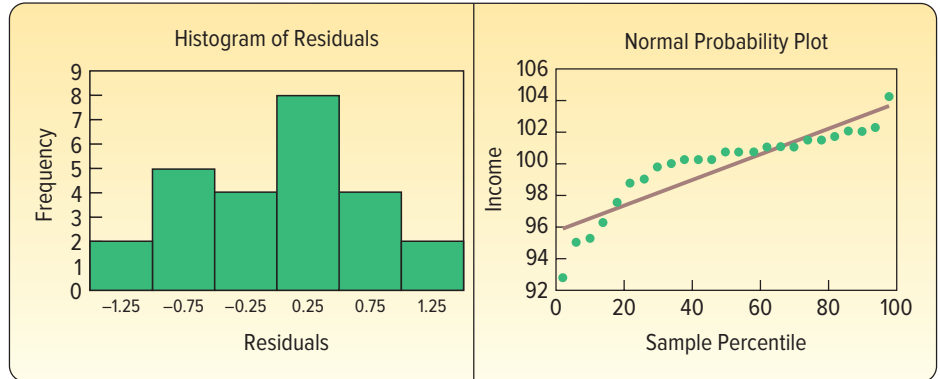
The scatter plot of income versus home value shows a general increasing trend. As the home value increases, so does family income. The points appear to be linear. That is, there is no observable nonlinear pattern in the data. The scatter plot on the right, of income versus years of education, shows that the data are measured to the nearest year. The measurement is to the nearest year and is a discrete variable. Given the measurement method, it is difficult to determine if the relationship is linear or not.

A plot of the residuals is also useful to evaluate the overall assumption of linearity. Recall that a residual is  $(y - \hat{y})$ , the difference between the actual value of the dependent variable ( $y$ ) and the predicted value of the dependent variable ( $\hat{y}$ ). Assuming a linear relationship, the distribution of the residuals should show about an equal proportion of negative residuals (points above the line) and positive residuals (points below the line), centered on zero. There should be no observable pattern to the plots. The graph follows.



There is no discernable pattern to the plot, so we conclude that the linearity assumption is reasonable.

If the linearity assumption is valid, then the distribution of residuals should follow the normal probability distribution with a mean of zero. To evaluate this assumption, we will use a histogram and a normal probability plot.



In general, the histogram on the left shows the major characteristics of a normal distribution, that is, a majority of observations in the middle and centered on the mean of zero, with lower frequencies in the tails of the distribution. The normal probability plot on the right is based on a cumulative normal probability distribution. The line shows the standardized cumulative normal distribution. The green dots show the cumulative distribution of the residuals. To confirm the normal distribution of the residuals, the green dots should be close to the line. This is true for most of the plot. However, we would note that there are departures and even perhaps a nonlinear pattern in the residuals in the lower part of the graph. As before, we are looking for serious departures from linearity and these are not indicated in these graphs.

The final assumption refers to multicollinearity. This means that the independent variables should not be highly correlated. We suggested a rule of thumb that multicollinearity would be a concern if the correlations among independent variables were close to 0.7 or -0.7. There are no violations of this guideline.

There is a statistic that is used to more precisely evaluate multicollinearity, the variance inflation factor (*VIF*). To calculate the *VIFs*, we need to do a regression analysis for each independent variable as a function of the other independent variables. From each of these regression analyses, we need the  $R^2$  to compute the *VIF* using formula (14–7). The following table shows the  $R^2$  for each regression analysis and the computed *VIF*. If the *VIFs* are less than 10, then multicollinearity is not a concern. In this case, the *VIFs* are all less than 10, so multicollinearity among the independent variables is not a concern.

| Dependent Variable | Independent Variables | $R^2$ -square | <i>VIF</i> |
|--------------------|-----------------------|---------------|------------|
| Value              | Education and Gender  | 0.058         | 1.062      |
| Education          | Gender and Value      | 0.029         | 1.030      |
| Gender             | Value and Education   | 0.042         | 1.044      |

To summarize, the multiple regression equation is

$$\hat{y} = 74.527 + 0.063(\text{Value}) + 1.016(\text{Education}) + 1.770(\text{Gender})$$

This equation explains 71.6% of the variation in family income. There are no major departures from the multiple regression assumptions of linearity, normally distributed residuals, and multicollinearity.

## CHAPTER SUMMARY

- I. The general form of a multiple regression equation is:

$$\hat{y} = a + b_1x_1 + b_2x_2 + \dots + b_kx_k \quad (14-1)$$

where  $a$  is the  $y$ -intercept when all  $x$ 's are zero,  $b_j$  refers to the sample regression coefficients, and  $x_j$  refers to the value of the various independent variables.

- A. There can be any number of independent variables.
  - B. The least squares criterion is used to develop the regression equation.
  - C. A statistical software package is needed to perform the calculations.
- II. An ANOVA table summarizes the multiple regression analysis.
- A. It reports the total amount of the variation in the dependent variable and divides this variation into that explained by the set of independent variables and that not explained.
  - B. It reports the degrees of freedom associated with the independent variables, the error variation, and the total variation.
- III. There are two measures of the effectiveness of the regression equation.
- A. The multiple standard error of estimate is similar to the standard deviation.
    1. It is measured in the same units as the dependent variable.
    2. It is based on squared deviations between the observed and predicted values of the dependent variable.
    3. It ranges from 0 to positive infinity.
    4. It is calculated from the following equation.

$$S_{y.123\dots k} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - (k + 1)}} \quad (14-2)$$

- B. The coefficient of multiple determination reports the percent of the variation in the dependent variable explained by the variation in the set of independent variables.
  1. It may range from 0 to 1.
  2. It is also based on squared deviations from the regression equation.
  3. It is found by the following equation.

$$R^2 = \frac{SSR}{SS \text{ total}} \quad (14-3)$$

- 4. When the number of independent variables is large, we adjust the coefficient of determination for the degrees of freedom as follows.

$$R_{\text{adj}}^2 = 1 - \frac{\frac{SSE}{n - (k + 1)}}{\frac{SS \text{ total}}{n - 1}} \quad (14-4)$$

- IV. A global test is used to investigate whether any of the independent variables have a regression coefficient that differs significantly from zero.
- A. The null hypothesis is: All the regression coefficients are zero.
  - B. The alternate hypothesis is: At least one regression coefficient is not zero.
  - C. The test statistic is the  $F$  distribution with  $k$  (the number of independent variables) degrees of freedom in the numerator and  $n - (k + 1)$  degrees of freedom in the denominator, where  $n$  is the sample size.
  - D. The formula to calculate the value of the test statistic for the global test is:

$$F = \frac{SSR/k}{SSE/[n - (k + 1)]} \quad (14-5)$$

- V. The test for individual variables determines which independent variables have regression coefficients that differ significantly from zero.
  - A. The variables that have zero regression coefficients are usually dropped from the analysis.
  - B. The test statistic is the  $t$  distribution with  $n - (k + 1)$  degrees of freedom.

C. The formula to calculate the value of the test statistic for the individual test is:

$$t = \frac{b_i - 0}{S_{b_i}} \tag{14-6}$$

VI. There are five assumptions to use multiple regression analysis.

- A. The relationship between the dependent variable and the set of independent variables must be linear.
  - 1. To verify this assumption, develop a scatter diagram and plot the residuals on the vertical axis and the fitted values on the horizontal axis.
  - 2. If the plots appear random, we conclude the relationship is linear.
- B. The variation is the same for both large and small values of  $\hat{y}$ .
  - 1. Homoscedasticity means the variation is the same for all fitted values of the dependent variable.
  - 2. This condition is checked by developing a scatter diagram with the residuals on the vertical axis and the fitted values on the horizontal axis.
  - 3. If there is no pattern to the plots—that is, they appear random—the residuals meet the homoscedasticity requirement.
- C. The residuals follow the normal probability distribution.
  - 1. This condition is checked by developing a histogram of the residuals or a normal probability plot.
  - 2. The mean of the distribution of the residuals is 0.
- D. The independent variables are not correlated.
  - 1. A correlation matrix of the independent variables with correlations larger than 0.70 or less than -0.70 indicates multicollinearity.
  - 2. Signs of correlated independent variables include when an important predictor variable is found to be insignificant, when an obvious reversal occurs in signs in one or more of the independent variables, or when a variable is removed from the solution and there is a large change in the regression coefficients.
  - 3. The variance inflation factor is used to identify correlated independent variables.

$$VIF = \frac{1}{1 - R_j^2} \tag{14-7}$$

- E. Each residual is independent of other residuals.
  - 1. Autocorrelation occurs when successive residuals are correlated.
  - 2. When autocorrelation exists, the value of the standard error will be biased and will return poor results for tests of hypothesis regarding the regression coefficients.

VII. Several techniques help build a regression model.

- A. A dummy or qualitative independent variable can assume one of two possible outcomes.
  - 1. A value of 1 is assigned to one outcome and 0 to the other.
  - 2. Use formula (14-6) to determine if the dummy variable should remain in the equation.
- B. Stepwise regression is a step-by-step process to find the regression equation.
  - 1. Only independent variables with nonzero regression coefficients enter the equation.
  - 2. Independent variables are added one at a time to the regression equation.

**PRONUNCIATION KEY**

| SYMBOL             | MEANING   | PRONUNCIATION                      |
|--------------------|---|------------------------------------|
| $b_1$              | Regression coefficient for the first independent variable | <i>b sub 1</i>                     |
| $b_k$              | Regression coefficient for any independent variable       | <i>b sub k</i>                     |
| $S_{y.123\dots k}$ | Multiple standard error of estimate                       | <i>s sub y dot 1, 2, 3 . . . k</i> |



## CHAPTER EXERCISES

11. A multiple regression analysis yields the following partial results.

| Source     | Sum of Squares | df |
|------------|----------------|----|
| Regression | 750            | 4  |
| Error      | 500            | 35 |

- What is the total sample size?
  - How many independent variables are being considered?
  - Compute the coefficient of determination.
  - Compute the standard error of estimate.
  - Test the hypothesis that at least one of the regression coefficients is not equal to zero. Let  $\alpha = .05$ .
12. In a multiple regression analysis, two independent variables are considered, and the sample size is 25. The regression coefficients and the standard errors are as follows.

$$\begin{aligned} b_1 &= 2.676 & s_{b_1} &= 0.56 \\ b_2 &= -0.880 & s_{b_2} &= 0.71 \end{aligned}$$

Conduct a test of hypothesis to determine whether either independent variable has a coefficient equal to zero. Would you consider deleting either variable from the regression equation? Use the .05 significance level.

13. The following output is from a multiple regression analysis.

| Analysis of Variance |    |     |    |
|----------------------|----|-----|----|
| Source               | DF | SS  | MS |
| Regression           | 5  | 100 | 20 |
| Residual Error       | 20 | 40  | 2  |
| Total                | 25 | 140 |    |

| Predictor | Coefficient | SE Coefficient | t     |
|-----------|-------------|----------------|-------|
| Constant  | 3.00        | 1.50           | 2.00  |
| $x_1$     | 4.00        | 3.00           | 1.33  |
| $x_2$     | 3.00        | 0.20           | 15.00 |
| $x_3$     | 0.20        | 0.05           | 4.00  |
| $x_4$     | -2.50       | 1.00           | -2.50 |
| $x_5$     | 3.00        | 4.00           | 0.75  |

- What is the sample size?
  - Compute the value of  $R^2$ .
  - Compute the multiple standard error of estimate.
  - Conduct a global test of hypothesis to determine whether any of the regression coefficients are significant. Use the .05 significance level.
  - Test the regression coefficients individually. Would you consider omitting any variable(s)? If so, which one(s)? Use the .05 significance level.
14. In a multiple regression analysis,  $k = 5$  and  $n = 20$ , the MSE value is 5.10, and SS total is 519.68. At the .05 significance level, can we conclude that any of the regression coefficients are not equal to 0?
15. The district manager of Jasons, a large discount electronics chain, is investigating why certain stores in her region are performing better than others. She believes that three factors are related to total sales: the number of competitors in the region, the population in the surrounding area, and the amount spent on advertising. From her district consisting of several hundred stores, she selects a random sample of 30 stores. For each store, she gathered the following information.
- $y$  = total sales last year (in \$ thousands)
  - $x_1$  = number of competitors in the region
  - $x_2$  = population of the region (in millions)
  - $x_3$  = advertising expense (in \$ thousands)

The results of a multiple regression analysis, using Minitab, follow.

| Analysis of Variance |    |      |         |
|----------------------|----|------|---------|
| Source               | DF | SS   | MS      |
| Regression           | 3  | 3050 | 1016.67 |
| Residual Error       | 26 | 2200 | 84.62   |
| Total                | 29 | 5250 |         |

| Predictor | Coefficient | SE Coefficient | t     |
|-----------|-------------|----------------|-------|
| Constant  | 14.00       | 7.00           | 2.00  |
| $x_1$     | -1.00       | 0.70           | -1.43 |
| $x_2$     | 30.00       | 5.20           | 5.77  |
| $x_3$     | 0.20        | 0.08           | 2.50  |

- a. What are the estimated sales for the Bryne store, which has four competitors, a regional population of 0.4 (400,000), and an advertising expense of 30 (\$30,000)?
  - b. Compute the  $R^2$  value.
  - c. Compute the multiple standard error of estimate.
  - d. Conduct a global test of hypothesis to determine whether any of the regression coefficients are not equal to zero. Use the .05 level of significance.
  - e. Conduct tests of hypothesis to determine which of the independent variables have significant regression coefficients. Which variables would you consider eliminating? Use the .05 significance level.
16. The sales manager of a large automotive parts distributor wants to estimate the total annual sales for each of the company's regions. Five factors appear to be related to regional sales: the number of retail outlets in the region, the number of automobiles in the region registered as of April 1, the total personal income recorded in the first quarter of the year, the average age of the automobiles (years), and the number of sales supervisors in the region. The data for each region were gathered for last year. For example, see the following table. In region 1 there were 1,739 retail outlets stocking the company's automotive parts, there were 9,270,000 registered automobiles in the region as of April 1, and so on. The region's sales for that year were \$37,702,000.

| Annual Sales (\$ millions), $y$ | Number of Retail Outlets, $x_1$ | Number of Automobiles Registered (millions), $x_2$ | Personal Income (\$ billions), $x_3$ | Average Age of Automobiles (years), $x_4$ | Number of Supervisors, $x_5$ |
|---------------------------------|---------------------------------|--|--------------------------------------|---|------------------------------|
| 37.702                          | 1,739                           | 9.27   | 85.4                                 | 3.5                                       | 9.0                          |
| 24.196                          | 1,221                           | 5.86   | 60.7                                 | 5.0                                       | 5.0                          |
| 32.055                          | 1,846                           | 8.81   | 68.1                                 | 4.4                                       | 7.0                          |
| 3.611                           | 120                             | 3.81   | 20.2                                 | 4.0                                       | 5.0                          |
| 17.625                          | 1,096                           | 10.31  | 33.8                                 | 3.5                                       | 7.0                          |
| 45.919                          | 2,290                           | 11.62  | 95.1                                 | 4.1                                       | 13.0                         |
| 29.600                          | 1,687                           | 8.96   | 69.3                                 | 4.1                                       | 15.0                         |
| 8.114                           | 241                             | 6.28   | 16.3                                 | 5.9                                       | 11.0                         |
| 20.116                          | 649                             | 7.77   | 34.9                                 | 5.5                                       | 16.0                         |
| 12.994                          | 1,427                           | 10.92  | 15.1                                 | 4.1                                       | 10.0                         |

- a. Consider the following correlation matrix. Which single variable has the strongest correlation with the dependent variable? The correlations between the independent variables outlets and income and between outlets and number of automobiles are fairly strong. Could this be a problem? What is this condition called?

|             |        |         |        |        |       |
|-------------|--------|---------|--------|--------|-------|
|             | sales  | outlets | cars   | income | age   |
| outlets     | 0.899  |         |        |        |       |
| automobiles | 0.605  | 0.775   |        |        |       |
| income      | 0.964  | 0.825   | 0.409  |        |       |
| age         | -0.323 | -0.489  | -0.447 | -0.349 |       |
| bosses      | 0.286  | 0.183   | 0.395  | 0.155  | 0.291 |

b. The output for all five variables is shown below. What percent of the variation is explained by the regression equation?

The regression equation is  
 $Sales = -19.7 - 0.00063 \text{ outlets} + 1.74 \text{ autos} + 0.410 \text{ income} + 2.04 \text{ age} - 0.034 \text{ bosses}$

| Predictor   | Coef      | SE Coef  | T     | P     |
|-------------|-----------|----------|-------|-------|
| Constant    | -19.672   | 5.422    | -3.63 | 0.022 |
| outlets     | -0.000629 | 0.002638 | -0.24 | 0.823 |
| automobiles | 1.7399    | 0.5530   | 3.15  | 0.035 |
| income      | 0.40994   | 0.04385  | 9.35  | 0.001 |
| age         | 2.0357    | 0.8779   | 2.32  | 0.081 |
| bosses      | -0.0344   | 0.1880   | -0.18 | 0.864 |

| Analysis of Variance |    |         |        |        |       |  |
|----------------------|----|---------|--------|--------|-------|--|
| SOURCE               | DF | SS      | MS     | F      | P     |  |
| Regression           | 5  | 1593.81 | 318.76 | 140.36 | 0.000 |  |
| Residual Error       | 4  | 9.08    | 2.27   |        |       |  |
| Total                | 9  | 1602.89 |        |        |       |  |

- c. Conduct a global test of hypothesis to determine whether any of the regression coefficients are not zero. Use the .05 significance level.
- d. Conduct a test of hypothesis on each of the independent variables. Would you consider eliminating “outlets” and “bosses”? Use the .05 significance level.
- e. The regression has been rerun below with “outlets” and “bosses” eliminated. Compute the coefficient of determination. How much has  $R^2$  changed from the previous analysis?

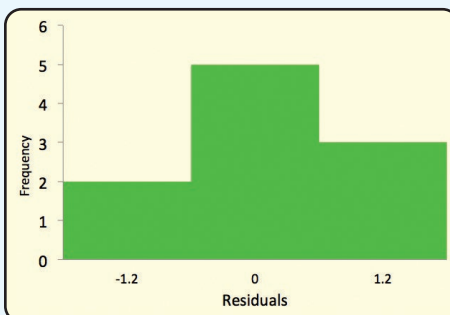
The regression equation is  
 $Sales = -18.9 + 1.61 \text{ autos} + 0.400 \text{ income} + 1.96 \text{ age}$

| Predictor   | Coef    | SE Coef | T     | P     |
|-------------|---------|---------|-------|-------|
| Constant    | -18.924 | 3.636   | -5.20 | 0.002 |
| automobiles | 1.6129  | 0.1979  | 8.15  | 0.000 |
| income      | 0.40031 | 0.01569 | 25.52 | 0.000 |
| age         | 1.9637  | 0.5846  | 3.36  | 0.015 |

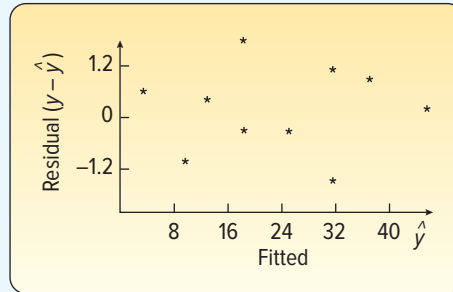
  

| Analysis of Variance |    |         |        |        |       |  |
|----------------------|----|---------|--------|--------|-------|--|
| SOURCE               | DF | SS      | MS     | F      | P     |  |
| Regression           | 3  | 1593.66 | 531.22 | 345.25 | 0.000 |  |
| Residual Error       | 6  | 9.23    | 1.54   |        |       |  |
| Total                | 9  | 1602.89 |        |        |       |  |

f. Following is a histogram of the residuals. Does the normality assumption appear reasonable? Why?



- g. Following is a plot of the fitted values of  $y$  (i.e.,  $\hat{y}$ ) and the residuals. What do you observe? Do you see any violations of the assumptions?



17. The administrator of a new paralegal program at Seagate Technical College wants to estimate the grade point average in the new program. He thought that high school GPA, the verbal score on the Scholastic Aptitude Test (SAT), and the mathematics score on the SAT would be good predictors of paralegal GPA. The data on nine students are:

| Student | High School GPA | SAT Verbal | SAT Math | Paralegal GPA |
|---------|-----------------|------------|----------|---------------|
| 1       | 3.25            | 480        | 410      | 3.21          |
| 2       | 1.80            | 290        | 270      | 1.68          |
| 3       | 2.89            | 420        | 410      | 3.58          |
| 4       | 3.81            | 500        | 600      | 3.92          |
| 5       | 3.13            | 500        | 490      | 3.00          |
| 6       | 2.81            | 430        | 460      | 2.82          |
| 7       | 2.20            | 320        | 490      | 1.65          |
| 8       | 2.14            | 530        | 480      | 2.30          |
| 9       | 2.63            | 469        | 440      | 2.33          |

- a. Consider the following correlation matrix. Which variable has the strongest correlation with the dependent variable? Some of the correlations among the independent variables are strong. Does this appear to be a problem?

|                 | Paralegal GPA | High School GPA | SAT Verbal |
|-----------------|---------------|-----------------|------------|
| High School GPA | 0.911         |                 |            |
| SAT Verbal      | 0.616         | 0.609           |            |
| SAT Math        | 0.487         | 0.636           | 0.599      |

- b. Consider the following output. Compute the coefficient of multiple determination.

The regression equation is  
 Paralegal GPA = -0.411 + 1.20 HSGPA + 0.00163 SAT\_Verbal - 0.00194 SAT\_Math

| Predictor  | Coef      | SE Coef  | T     | P     |
|------------|-----------|----------|-------|-------|
| Constant   | -0.4111   | 0.7823   | -0.53 | 0.622 |
| HSGPA      | 1.2014    | 0.2955   | 4.07  | 0.010 |
| SAT_Verbal | 0.001629  | 0.002147 | 0.76  | 0.482 |
| SAT_Math   | -0.001939 | 0.002074 | -0.94 | 0.393 |

| Analysis of Variance |    |        |        |       |       |
|----------------------|----|--------|--------|-------|-------|
| SOURCE               | DF | SS     | MS     | F     | P     |
| Regression           | 3  | 4.3595 | 1.4532 | 10.33 | 0.014 |
| Residual Error       | 5  | 0.7036 | 0.1407 |       |       |
| Total                | 8  | 5.0631 |        |       |       |

| SOURCE     | DF | Seq SS |
|------------|----|--------|
| HSGPA      | 1  | 4.2061 |
| SAT_Verbal | 1  | 0.0303 |
| SAT_Math   | 1  | 0.1231 |

- c. Conduct a global test of hypothesis from the preceding output. Does it appear that any of the regression coefficients are not equal to zero?
- d. Conduct a test of hypothesis on each independent variable. Would you consider eliminating the variables “SAT\_Verbal” and “SAT\_Math”? Let  $\alpha = .05$ .
- e. The analysis has been rerun without “SAT\_Verbal” and “SAT\_Math.” See the following output. Compute the coefficient of determination. How much has  $R^2$  changed from the previous analysis?

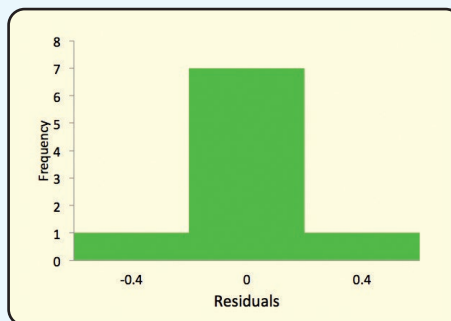
The regression equation is  
Paralegal GPA =  $-0.454 + 1.16$  HSGPA

| Predictor | Coef    | SE Coef | T     | P     |
|-----------|---------|---------|-------|-------|
| Constant  | -0.4542 | 0.5542  | -0.82 | 0.439 |
| HSGPA     | 1.1589  | 0.1977  | 5.86  | 0.001 |

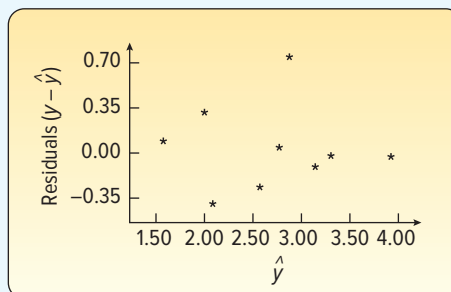
  

| Analysis of Variance |    |        |        |       |       |
|----------------------|----|--------|--------|-------|-------|
| SOURCE               | DF | SS     | MS     | F     | P     |
| Regression           | 1  | 4.2061 | 4.2061 | 34.35 | 0.001 |
| Residual Error       | 7  | 0.8570 | 0.1224 |       |       |
| Total                | 8  | 5.0631 |        |       |       |

- f. Following is a histogram of the residuals. Does the normality assumption for the residuals seem reasonable?



- g. Following is a plot of the residuals and the  $\hat{y}$  values. Do you see any violation of the assumptions?



18. **FILE** Mike Wilde is president of the teachers' union for Otsego School District. In preparing for upcoming negotiations, he is investigating the salary structure of classroom teachers in the district. He believes there are three factors that affect a teacher's salary: years of experience, a teaching effectiveness rating given by the principal, and whether the teacher has a master's degree. A random sample of 20 teachers resulted in the following data.

| Salary<br>(\$ thousands),<br>$y$ | Years of<br>Experience,<br>$x_1$ | Principal's<br>Rating,<br>$x_2$ | Master's<br>Degree,*<br>$x_3$ |
|----------------------------------|----------------------------------|---------------------------------|-------------------------------|
| 31.1                             | 8                                | 35                              | 0                             |
| 33.6                             | 5                                | 43                              | 0                             |
| 29.3                             | 2                                | 51                              | 1                             |
| ⋮                                | ⋮                                | ⋮                               | ⋮                             |
| 30.7                             | 4                                | 62                              | 0                             |
| 32.8                             | 2                                | 80                              | 1                             |
| 42.8                             | 8                                | 72                              | 0                             |

\*1 = yes, 0 = no.

- Develop a correlation matrix. Which independent variable has the strongest correlation with the dependent variable? Does it appear there will be any problems with multicollinearity?
  - Determine the regression equation. What salary would you estimate for a teacher with 5 years' experience, a rating by the principal of 60, and no master's degree?
  - Conduct a global test of hypothesis to determine whether any of the regression coefficients differ from zero. Use the .05 significance level.
  - Conduct a test of hypothesis for the individual regression coefficients. Would you consider deleting any of the independent variables? Use the .05 significance level.
  - If your conclusion in part (d) was to delete one or more independent variables, run the analysis again without those variables.
  - Determine the residuals for the equation of part (e). Use a histogram to verify that the distribution of the residuals is approximately normal.
  - Plot the residuals computed in part (f) in a scatter diagram with the residuals on the Y-axis and the  $\hat{y}$  values on the X-axis. Does the plot reveal any violations of the assumptions of regression?
19. **FILE** A video media consultant collected the following data on popular LED televisions sold through online retailers.

| Manufacturer | Screen | Price   | Manufacturer | Screen | Price   |
|--------------|--------|---------|--------------|--------|---------|
| Sharp        | 46     | 736.50  | Sharp        | 37     | 657.25  |
| Samsung      | 52     | 1150.00 | Sharp        | 32     | 426.75  |
| Samsung      | 46     | 895.00  | Sharp        | 52     | 1389.00 |
| Sony         | 40     | 625.00  | Samsung      | 40     | 874.75  |
| Sharp        | 42     | 773.25  | Sharp        | 32     | 517.50  |
| Samsung      | 46     | 961.25  | Samsung      | 52     | 1475.00 |
| Samsung      | 40     | 686.00  | Sony         | 40     | 954.25  |
| Sharp        | 37     | 574.75  | Sony         | 52     | 1551.50 |
| Sharp        | 46     | 1000.00 | Sony         | 46     | 1303.00 |
| Sony         | 40     | 722.25  | Sony         | 46     | 1430.50 |
| Sony         | 52     | 1307.50 | Sony         | 52     | 1717.00 |
| Samsung      | 32     | 373.75  |              |        |         |

- Does there appear to be a linear relationship between the screen size and the price?
- Which variable is the "dependent" variable?
- Using statistical software, determine the regression equation. Interpret the value of the slope in the regression equation.
- Include the manufacturer in a multiple linear regression analysis using a "dummy" variable. Does it appear that some manufacturers can command a premium price? Hint: You will need to use a set of dummy or indicator variables.
- Test each of the individual coefficients to see if they are significant.
- Make a plot of the residuals and comment on whether they appear to follow a normal distribution.
- Plot the residuals versus the fitted values. Do they seem to have the same amount of variation?

20. **FILE** A regional planner is studying the demographics of nine counties in the eastern region of an Atlantic seaboard state. She has gathered the following data:

| County | Median Income | Median Age | Coastal* |
|--------|---------------|------------|----------|
| A      | \$48,157      | 57.7       | 1        |
| B      | 48,568        | 60.7       | 1        |
| C      | 46,816        | 47.9       | 1        |
| D      | 34,876        | 38.4       | 0        |
| E      | 35,478        | 42.8       | 0        |
| F      | 34,465        | 35.4       | 0        |
| G      | 35,026        | 39.5       | 0        |
| H      | 38,599        | 65.6       | 0        |
| J      | 33,315        | 27.0       | 0        |

\*1 = yes, 0 = no.

- Is there a linear relationship between the median income and median age?
  - Which variable is the “dependent” variable?
  - Use statistical software to determine the regression equation. Interpret the value of the slope in a simple regression equation.
  - Include the aspect that the county is “coastal” or not in a multiple linear regression analysis using a “dummy” variable. Does it appear to be a significant influence on incomes?
  - Test each of the individual coefficients to see if they are significant.
  - Make a plot of the residuals and comment on whether they appear to follow a normal distribution.
  - Plot the residuals versus the fitted values. Do they seem to have the same amount of variation?
21. **FILE** Great Plains Distributors Inc. sells roofing and siding products to home improvement retailers, such as Lowe’s and Home Depot, and commercial contractors. The owner is interested in studying the effects of several variables on the sales volume of fiber-cement siding products.

The company has 26 marketing districts across the United States. In each district, it collected information on the following variables: sales volume (in thousands of dollars), advertising dollars (in thousands), number of active accounts, number of competing brands, and a rating of market potential.

| Sales<br>(000s) | Advertising       |                       |                          | Market<br>Potential |
|-----------------|-------------------|-----------------------|--------------------------|---------------------|
|                 | Dollars<br>(000s) | Number of<br>Accounts | Number of<br>Competitors |                     |
| 79.3            | 5.5               | 31                    | 10                       | 8                   |
| 200.1           | 2.5               | 55                    | 8                        | 6                   |
| 163.2           | 8.0               | 67                    | 12                       | 9                   |
| 200.1           | 3.0               | 50                    | 7                        | 16                  |
| 146.0           | 3.0               | 38                    | 8                        | 15                  |
| 177.7           | 2.9               | 71                    | 12                       | 17                  |
| ⋮               | ⋮                 | ⋮                     | ⋮                        | ⋮                   |
| 93.5            | 4.2               | 26                    | 8                        | 3                   |
| 259.0           | 4.5               | 75                    | 8                        | 19                  |
| 331.2           | 5.6               | 71                    | 4                        | 9                   |

Conduct a multiple regression analysis to find the best predictors of sales.

- Draw a scatter diagram comparing sales volume with each of the independent variables. Comment on the results.
- Develop a correlation matrix. Do you see any problems? Does it appear there are any redundant independent variables?

- c. Develop a regression equation. Conduct the global test. Can we conclude that some of the independent variables are useful in explaining the variation in the dependent variable?
  - d. Conduct a test of each of the independent variables. Are there any that should be dropped?
  - e. Refine the regression equation so the remaining variables are all significant.
  - f. Develop a histogram of the residuals and a normal probability plot. Are there any problems?
  - g. Determine the variance inflation factor for each of the independent variables. Are there any problems?
22. **FILE** A market researcher is studying online subscription services. She is particularly interested in what variables relate to the number of subscriptions for a particular online service. She is able to obtain the following sample information on 25 online subscription services. The following notation is used:

Sub = Number of subscriptions (in thousands)  
 Web page hits = Average monthly count (in thousands)  
 Adv = Advertising budget of the service (in \$ hundreds)  
 Price = Average monthly subscription price (\$)

| Service | Web Page |       |      |       |
|---------|----------|-------|------|-------|
|         | Sub      | Hits  | Adv  | Price |
| 1       | 37.95    | 588.9 | 13.2 | 35.1  |
| 2       | 37.66    | 585.3 | 13.2 | 34.7  |
| 3       | 37.55    | 566.3 | 19.8 | 34.8  |
| ⋮       | ⋮        | ⋮     | ⋮    | ⋮     |
| 23      | 38.83    | 629.6 | 22.0 | 35.3  |
| 24      | 38.33    | 680.0 | 24.2 | 34.7  |
| 25      | 40.24    | 651.2 | 33.0 | 35.8  |

- a. Determine the regression equation.
  - b. Conduct a global test of hypothesis to determine whether any of the regression coefficients are not equal to zero.
  - c. Conduct a test for the individual coefficients. Would you consider deleting any coefficients?
  - d. Determine the residuals and plot them against the fitted values. Do you see any problems?
  - e. Develop a histogram of the residuals. Do you see any problems with the normality assumption?
23. **FILE** Fred G. Hire is the manager of human resources at Crescent Custom Steel Products. As part of his yearly report to the CEO, he is required to present an analysis of the salaried employees. For each of the 30 salaried employees, he records monthly salary; service at Crescent, in months; age; gender (1 = male, 0 = female); and whether the employee has a management or engineering position. Those employed in management are coded 0, and those in engineering are coded 1.

| Sampled Employee | Monthly Salary | Length of Service | Age | Gender | Job |
|------------------|----------------|-------------------|-----|--------|-----|
| 1                | \$1,769        | 93                | 42  | 1      | 0   |
| 2                | 1,740          | 104               | 33  | 1      | 0   |
| 3                | 1,941          | 104               | 42  | 1      | 1   |
| ⋮                | ⋮              | ⋮                 | ⋮   | ⋮      | ⋮   |
| 28               | 1,791          | 131               | 56  | 0      | 1   |
| 29               | 2,001          | 95                | 30  | 1      | 1   |
| 30               | 1,874          | 98                | 47  | 1      | 0   |



- a. Determine the regression equation, using salary as the dependent variable and the other four variables as independent variables.
  - b. What is the value of  $R^2$ ? Comment on this value.
  - c. Conduct a global test of hypothesis to determine whether any of the independent variables are different from 0.
  - d. Conduct an individual test to determine whether any of the independent variables can be dropped.
  - e. Rerun the regression equation, using only the independent variables that are significant. How much more does a man earn per month than a woman? Does it make a difference whether the employee has a management or engineering position?
24. **FILE** Many regions in North Carolina, South Carolina, and Georgia have experienced rapid population growth over the last 10 years. It is expected that the growth will continue over the next 10 years. This has motivated many of the large grocery store chains to build new stores in the region. The Kelley's Super Grocery Stores Inc. chain is no exception. The director of planning for Kelley's Super Grocery Stores wants to study adding more stores in this region. He believes there are two main factors that indicate the amount families spend on groceries. The first is their income and the other is the number of people in the family. The director gathered the following sample information.

| Family | Food   | Income  | Size |
|--------|--------|---------|------|
| 1      | \$5.04 | \$73.98 | 4    |
| 2      | 4.08   | 54.90   | 2    |
| 3      | 5.76   | 94.14   | 4    |
| ⋮      | ⋮      | ⋮       | ⋮    |
| 23     | 4.56   | 38.16   | 3    |
| 24     | 5.40   | 43.74   | 7    |
| 25     | 4.80   | 48.42   | 5    |

Food and income are reported in thousands of dollars per year, and the variable size refers to the number of people in the household.

- a. Develop a correlation matrix. Do you see any problems with multicollinearity?
  - b. Determine the regression equation. Discuss the regression equation. How much does an additional family member add to the amount spent on food?
  - c. What is the value of  $R^2$ ? Can we conclude that this value is greater than 0?
  - d. Would you consider deleting either of the independent variables?
  - e. Plot the residuals in a histogram. Is there any problem with the normality assumption?
  - f. Plot the fitted values against the residuals. Does this plot indicate any problems with homoscedasticity?
25. **FILE** An investment advisor is studying the relationship between a common stock's price to earnings (P/E) ratio and factors that she thinks would influence it. She has the following data on the earnings per share (EPS) and the dividend percentage (Yield) for a sample of 20 stocks.

| Stock | P/E   | EPS    | Yield |
|-------|-------|--------|-------|
| 1     | 20.79 | \$2.46 | 1.42  |
| 2     | 3.03  | 2.69   | 4.05  |
| 3     | 44.46 | -0.28  | 4.16  |
| ⋮     | ⋮     | ⋮      | ⋮     |
| 18    | 30.21 | 1.71   | 3.07  |
| 19    | 32.88 | 0.35   | 2.21  |
| 20    | 15.19 | 5.02   | 3.50  |

- a. Develop a multiple linear regression with P/E as the dependent variable.
- b. Are either of the two independent variables an effective predictor of P/E?
- c. Interpret the regression coefficients.
- d. Do any of these stocks look particularly undervalued?

- e. Plot the residuals and check the normality assumption. Plot the fitted values against the residuals.
  - f. Do there appear to be any problems with homoscedasticity?
  - g. Develop a correlation matrix. Do any of the correlations indicate multicollinearity?
26. **FILE** The Conch Café, located in Gulf Shores, Alabama, features casual lunches with a great view of the Gulf of Mexico. To accommodate the increase in business during the summer vacation season, Fuzzy Conch, the owner, hires a large number of servers as seasonal help. When he interviews a prospective server, he would like to provide data on the amount a server can earn in tips. He believes that the amount of the bill and the number of diners are both related to the amount of the tip. He gathered the following sample information.

| Customer | Amount of Tip | Amount of Bill | Number of Diners |
|----------|---------------|----------------|------------------|
| 1        | \$7.00        | \$48.97        | 5                |
| 2        | 4.50          | 28.23          | 4                |
| 3        | 1.00          | 10.65          | 1                |
| ⋮        | ⋮             | ⋮              | ⋮                |
| 28       | 2.50          | 26.25          | 2                |
| 29       | 9.25          | 56.81          | 5                |
| 30       | 8.25          | 50.65          | 5                |

- a. Develop a multiple regression equation with the amount of tips as the dependent variable and the amount of the bill and the number of diners as independent variables. Write out the regression equation. How much does another diner add to the amount of the tips?
  - b. Conduct a global test of hypothesis to determine if at least one of the independent variables is significant. What is your conclusion?
  - c. Conduct an individual test on each of the variables. Should one or the other be deleted?
  - d. Use the equation developed in part (c) to determine the coefficient of determination. Interpret the value.
  - e. Plot the residuals. Is it reasonable to assume they follow the normal distribution?
  - f. Plot the residuals against the fitted values. Is it reasonable to conclude they are random?
27. **FILE** The president of Blitz Sales Enterprises sells kitchen products through cable television infomercials. He gathered data from the last 15 weeks of sales to determine the relationship between sales and the number of infomercials.

| Infomercials | Sales (\$000s) | Infomercials | Sales (\$000s) |
|--------------|----------------|--------------|----------------|
| 20           | 3.2            | 22           | 2.5            |
| 15           | 2.6            | 15           | 2.4            |
| 25           | 3.4            | 25           | 3.0            |
| 10           | 1.8            | 16           | 2.7            |
| 18           | 2.2            | 12           | 2.0            |
| 18           | 2.4            | 20           | 2.6            |
| 15           | 2.4            | 25           | 2.8            |
| 12           | 1.5            |              |                |

- a. Determine the regression equation. Are the sales predictable from the number of commercials?
- b. Determine the residuals and plot a histogram. Does the normality assumption seem reasonable?

28. **FILE** The director of special events for Sun City believed that the amount of money spent on fireworks displays for the 4th of July was predictive of attendance at the Fall Festival held in October. She gathered the following data to test her suspicion.

| 4th of July (\$000) | Fall Festival (000) | 4th of July (\$000) | Fall Festival (000) |
|---------------------|---------------------|---------------------|---------------------|
| 10.6                | 8.8                 | 9.0                 | 9.5                 |
| 8.5                 | 6.4                 | 10.0                | 9.8                 |
| 12.5                | 10.8                | 7.5                 | 6.6                 |
| 9.0                 | 10.2                | 10.0                | 10.1                |
| 5.5                 | 6.0                 | 6.0                 | 6.1                 |
| 12.0                | 11.1                | 12.0                | 11.3                |
| 8.0                 | 7.5                 | 10.5                | 8.8                 |
| 7.5                 | 8.4                 |                     |                     |

Determine the regression equation. Is the amount spent on fireworks related to attendance at the Fall Festival? Evaluate the regression assumptions by examining the residuals.

## DATA ANALYTICS

(The data for these exercises are available in Connect.)

29. The North Valley Real Estate data report information on homes on the market. Use the selling price of the home as the dependent variable and determine the regression equation using the size of the house, number of bedrooms, days on the market, and number of bathrooms as independent variables.
- Develop a correlation matrix. Which independent variables have strong or weak correlations with the dependent variable? Do you see any problems with multicollinearity?
  - Use a statistical software package to determine the multiple regression equation. How did you select the variables to include in the equation? How did you use the information from the correlation analysis? Show that your regression equation shows a significant relationship. Write out the regression equation and interpret its practical application. Report and interpret the  $R$ -square.
  - Using your results from part (b), evaluate the addition of the variables pool or garage. Report your results and conclusions.
  - Develop a histogram of the residuals from the final regression equation developed in part (c). Is it reasonable to conclude that the normality assumption has been met?
  - Plot the residuals against the fitted values from the final regression equation developed in part (c). Plot the residuals on the vertical axis and the fitted values on the horizontal axis.
30. Refer to the Baseball 2016 data, which report information on the 30 Major League Baseball teams for the 2016 season. Let the number of games won be the dependent variable and the following variables be independent variables: team batting average, team earned run average (ERA), number of home runs, and whether the team plays in the American or the National League.
- Develop a correlation matrix. Which independent variables have strong or weak correlations with the dependent variable? Do you see any problems with multicollinearity? Are you surprised that the correlation coefficient for ERA is negative?
  - Use a statistical software package to determine the multiple regression equation. How did you select the variables to include in the equation? How did you use the

- information from the correlation analysis? Show that your regression equation shows a significant relationship. Write out the regression equation and interpret its practical application. Report and interpret the  $R$ -square. Is the number of wins affected by whether the team plays in the National or the American League?
- c. Conduct a global test on the set of independent variables. Interpret.
  - d. Conduct a test of hypothesis on each of the independent variables. Would you consider deleting any of the variables? If so, which ones?
  - e. Develop a histogram of the residuals from the final regression equation developed in part (d). Is it reasonable to conclude that the normality assumption has been met?
  - f. Plot the residuals against the fitted values from the final regression equation developed in part (d). Plot the residuals on the vertical axis and the fitted values on the horizontal axis.
- 31.** Refer to the Lincolnville School District bus data. First, add a variable to change the type of engine (diesel or gasoline) to a qualitative variable. If the engine type is diesel, then set the qualitative variable to 0. If the engine type is gasoline, then set the qualitative variable to 1. Develop a regression equation using statistical software with maintenance cost as the dependent variable and age, odometer miles, miles since last maintenance, and engine type as the independent variables.
- a. Develop a correlation matrix. Which independent variables have strong or weak correlations with the dependent variable? Do you see any problems with multicollinearity?
  - b. Use a statistical software package to determine the multiple regression equation. How did you select the variables to include in the equation? How did you use the information from the correlation analysis? Show that your regression equation shows a significant relationship. Write out the regression equation and interpret its practical application. Report and interpret the  $R$ -square.
  - c. Develop a histogram of the residuals from the final regression equation developed in part (b). Is it reasonable to conclude that the normality assumption has been met?
  - d. Plot the residuals against the fitted values from the final regression equation developed in part (b) against the fitted values of  $Y$ . Plot the residuals on the vertical axis and the fitted values on the horizontal axis.

## PRACTICE TEST

### Part 1—Objective

1. Multiple regression analysis describes the relationship between one dependent variable and two or more \_\_\_\_\_.
2. In multiple regression analysis, the regression coefficients are computed using the method of \_\_\_\_\_. (residuals, normality, least squares, standardization)
3. In multiple regression analysis, the multiple standard error of the estimate is the square root of the \_\_\_\_\_. (mean square error, residual, residual squared, explained variation)
4. The coefficient of multiple determination is the percent of variation in the dependent variable that is explained by the set of \_\_\_\_\_.
5. The adjusted coefficient of determination compensates for the number of \_\_\_\_\_. (dependent variables, errors, independent variables)
6. In the global test of the regression coefficients, when the hypothesis is rejected, at least one coefficient is \_\_\_\_\_.
7. The test statistic for the global test of regression coefficients is the \_\_\_\_\_.
8. The test statistic for testing individual regression coefficients is the \_\_\_\_\_.
9. A scatter plot of the residuals versus the fitted values of the dependent variable evaluates the assumption of \_\_\_\_\_.
10. Multicollinearity exists when independent variables are \_\_\_\_\_.
11. The variance inflation factor is used to detect \_\_\_\_\_.
12. Another term for a qualitative variable is a \_\_\_\_\_ variable.

### Part 2—Problems

1. Given the following ANOVA output:

| Source     | Sum of Squares | df | MS     |
|------------|----------------|----|--------|
| Regression | 1050.8         | 4  | 262.70 |
| Error      | 83.8           | 20 | 4.19   |
| Total      | 1134.6         | 24 |        |

| Predictor | Coefficient | St. Dev | t-ratio |
|-----------|-------------|---------|---------|
| Constant  | 70.06       | 2.13    | 32.89   |
| $X_1$     | 0.42        | 0.17    | 2.47    |
| $X_2$     | 0.27        | 0.21    | 1.29    |
| $X_3$     | 0.75        | 0.30    | 2.50    |
| $X_4$     | 0.42        | 0.07    | 6.00    |

- How many independent variables are there in the regression equation?
- Write out the regression equation.
- Compute the coefficient of multiple determination.
- Compute the multiple standard error of estimate.
- Conduct a hypothesis test to determine if any of the regression coefficients are different from zero.
- Conduct a hypothesis test on each of the regression coefficients. Can any of them be deleted?

# Nonparametric Methods:

# 15

## NOMINAL-LEVEL HYPOTHESIS TESTS



©Digital Vision/SuperStock

- ▲ **FOR MANY YEARS**, TV executives used the guideline that 30% of the audience were watching each of the traditional big three prime-time networks and 10% were watching cable stations on a weekday night. A random sample of 500 viewers in the Tampa–St. Petersburg, Florida, area last Monday night showed that 165 homes were tuned in to the ABC affiliate, 140 to the CBS affiliate, 125 to the NBC affiliate, and the remainder were viewing a cable station. At the .05 significance level, can we conclude that the guideline is still reasonable? (See Exercise 24 and [LO15-3](#).)

### LEARNING OBJECTIVES

---

*When you have completed this chapter, you will be able to:*

- LO15-1** Test a hypothesis about a population proportion.
- LO15-2** Test a hypothesis about two population proportions.
- LO15-3** Test a hypothesis comparing an observed set of frequencies to an expected frequency distribution.
- LO15-4** Explain the limitations of using the chi-square statistic in goodness-of-fit tests.
- LO15-5** Perform a chi-square test for independence on a contingency table.

## INTRODUCTION

In Chapters 9 through 12, we described tests of hypothesis for data of interval or ratio scale. Examples of interval- and ratio-scale data include the scores on the first statistics examination in your class, the incomes of corporate executive officers in technology companies, or years of employment of production workers at the BMW plant in Greer, South Carolina.

We conducted hypothesis tests about a single population mean (Chapter 10), about two population means (Chapter 11), and about three or more population means (Chapter 12). For these tests, we use interval or ratio data and assume the populations follow the normal probability distribution. However, there are hypothesis tests that do not require any assumption regarding the shape of the population. Hence, the assumption of a normal population is not necessary. These tests are referred to as nonparametric hypothesis tests.

In this chapter, we begin with tests of hypothesis for nominal-scale data. Recall that nominal-scale data are simply classified into mutually exclusive categories. In the first two sections of this chapter, we describe tests of proportions. In these tests, individuals or objects are classified into one of two mutually exclusive groups. Examples include job placement (employed or not employed), quality (acceptable or unacceptable), diabetes (yes or no), and airline flight arrivals (on time or late).

We also expand the nominal-scale tests to include situations where data are classified into several mutually exclusive categories. The scale of measurement is still nominal, but there are several categories. Examples include the colors of M&M Plain candies (red, green, blue, yellow, orange, and brown), brand of peanut butter purchased (Peter Pan, Jif, Skippy, and others), or days of the workweek (Monday, Tuesday, Wednesday, Thursday, and Friday). We introduce the chi-square distribution as a new test statistic. It is most often used when there are more than two nominal-scale categories.

### LO15-1

Test a hypothesis about a population proportion.

## TEST A HYPOTHESIS OF A POPULATION PROPORTION

Beginning on page 260 in Chapter 9, we discussed confidence intervals for proportions. We can also conduct a test of hypothesis for a proportion. Recall that a proportion is the ratio of the number of successes to the number of observations. We let  $X$  refer to the number of successes and  $n$  the number of observations, so the proportion of successes in a fixed number of trials is  $X/n$ . Thus, the formula for computing a sample proportion,  $p$ , is  $p = X/n$ . Consider the following potential hypothesis-testing situations.

- Historically, General Motors reports that 70% of leased vehicles are returned with less than 36,000 miles. A recent sample of 200 vehicles returned at the end of their lease showed 158 had less than 36,000 miles. Has the proportion increased?
- The American Association of Retired Persons (AARP) reports that 60% of retired people under the age of 65 would return to work on a full-time basis if a suitable job were available. A sample of 500 retirees under 65 revealed 315 would return to work. Can we conclude that more than 60% would return to work?
- Able Moving and Storage Inc. advises its clients for long-distance residential moves that their household goods will be delivered in 3 to 5 days from the time they are picked up. Able's records show it is successful 90% of the time with this claim. A recent audit revealed it was successful 190 times out of 200. Can the company conclude its success rate has increased?

Some assumptions must be made and conditions met before testing a population proportion. To test a hypothesis about a population proportion, a random sample is

chosen from the population. It is assumed that the binomial assumptions discussed in Chapter 6 are met: (1) the sample data collected are the result of counts; (2) the outcome of an experiment is classified into one of two mutually exclusive categories—a “success” or a “failure”; (3) the probability of a success is the same for each trial; and (4) the trials are independent, meaning the outcome of one trial does not affect the outcome of any other trial. This test is appropriate when both  $n\pi$  and  $n(1 - \pi)$  are at least 5.  $n$  is the sample size, and  $\pi$  is the population proportion. It takes advantage of the fact that a binomial distribution can be approximated by the normal distribution.

### EXAMPLE

A Republican governor of a western state is thinking about running for reelection. Historically, to be reelected, a Republican candidate needs to earn at least 80% of the vote in the northern section of the state. The governor hires a polling organization to survey the voters in the northern section of the state and determine what percent would vote for him. The polling organization will survey 2,000 voters. Use a statistical hypothesis-testing procedure to assess the governor’s chances of reelection.

### SOLUTION

This situation regarding the governor’s reelection meets the binomial conditions.

- There are only two possible outcomes. That is, a sampled voter will either vote or not vote for the governor.
- The probability of a success is the same for each trial. In this case, the likelihood a particular sampled voter will support reelection is .80.
- The trials are independent. This means, for example, the likelihood the 23rd voter sampled will support reelection is not affected by what the 24th or 52nd voter does.
- The sample data are the result of counts. We are going to count the number of voters who support reelection in the sample of 2,000.

We can use a normal approximation to the binomial distribution if both  $n\pi$  and  $n(1 - \pi)$  exceed 5. In this case,  $n = 2,000$  and  $\pi = 0.80$ . ( $\pi$  is the proportion of the vote in the northern part of the state, or 80%, needed to be elected.) Thus,  $n\pi = 2,000(.80) = 1,600$  and  $n(1 - \pi) = 2,000(1 - .80) = 400$ . Both 1,600 and 400 are clearly greater than 5.

**Step 1: State the null hypothesis and the alternate hypothesis.** The null hypothesis,  $H_0$ , is that the population proportion  $\pi$  is .80 or larger. The alternate hypothesis,  $H_1$ , is that the proportion is less than .80. From a practical standpoint, the incumbent governor is concerned only when the proportion is less than .80. If it is equal to or greater than .80, he will have no problem; that is, the sample data would indicate he will be reelected. These hypotheses are written symbolically as:

$$H_0: \pi \geq .80$$

$$H_1: \pi < .80$$

$H_1$  states a direction. Thus, as noted previously, the test is one-tailed with the inequality sign pointing to the tail of the distribution containing the region of rejection.

**Step 2: Select the level of significance.** The level of significance is .05. This is the likelihood that a true hypothesis will be rejected.



**Step 3: Select the test statistic.**  $z$  is the appropriate statistic, found by:

**TEST OF HYPOTHESIS,  
ONE PROPORTION**

$$z = \frac{\rho - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} \quad (15-1)$$

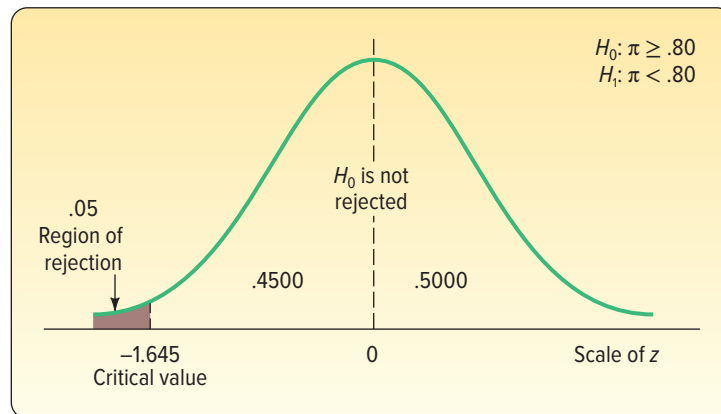
where:

$\pi$  is the population proportion.

$\rho$  is the sample proportion.

$n$  is the sample size.

**Step 4: Formulate the decision rule.** The critical value or values of  $z$  form the dividing point or points between the regions where  $H_0$  is rejected and where it is not rejected. Because the alternate hypothesis states a direction, this is a one-tailed test. The sign of the inequality points to the left, so only the left side of the curve is used. (See Chart 15–1.) The significance level is .05. This probability is in the left tail and determines the region of rejection. The area between zero and the critical value is .4500, found by  $.5000 - .0500$ . Referring to Appendix B.3, go to the column indicating a .05 significance level for a one-tailed test, find the row with infinite degrees of freedom, and read the  $z$  value of 1.645. The decision rule is, therefore: Reject the null hypothesis and accept the alternate hypothesis if the computed value of  $z$  falls to the left of  $-1.645$ ; otherwise do not reject  $H_0$ .



**CHART 15–1** Rejection Region for the .05 Level of Significance, One-Tailed Test

**Step 5: Make a decision.** Select a sample and make a decision about  $H_0$ . A sample survey of 2,000 potential voters in the northern part of the state revealed that 1,550 planned to vote for the incumbent governor. Is the sample proportion of .775 (found by  $1,550/2,000$ ) close enough to .80 to conclude that the difference is due to sampling error? In this case:

$\rho$  is .775, the proportion in the sample who plan to vote for the governor.

$n$  is 2,000, the number of voters surveyed.

$\pi$  is .80, the hypothesized population proportion.

$z$  is a normally distributed test statistic. We can use it because the normal approximation assumptions are true.

Using formula (15–1) and computing  $z$  gives

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} = \frac{\frac{1,550}{2,000} - .80}{\sqrt{\frac{.80(1 - .80)}{2,000}}} = \frac{.775 - .80}{\sqrt{.00008}} = -2.80$$

The computed value of  $z$  ( $-2.80$ ) is less than the critical value, so the null hypothesis is rejected at the .05 level. The difference of 2.5 percentage points between the sample percent (77.5%) and the hypothesized population percent in the northern part of the state necessary to carry the state (80%) is statistically significant. From Appendix B.3, the probability of a  $z$  value between zero and  $-2.80$  is .4974. So the  $p$ -value is .0026, found by  $.5000 - .4974$ . Because the  $p$ -value is less than the significance level, the null hypothesis is rejected.

**Step 6: Interpret the result.** The governor can conclude that he does not have the necessary support in the northern section of the state to win reelection. To put it another way, the evidence at this point does not support the claim that the incumbent governor will return to the governor's mansion for another 4 years.

## SELF-REVIEW 15–1



A recent insurance industry report indicated that 40% of those persons involved in minor traffic accidents this year have been involved in at least one other traffic accident in the last 5 years. An advisory group decided to investigate this claim, believing it was too large. A sample of 200 traffic accidents this year showed 74 persons were also involved in another accident within the last 5 years. Use the .01 significance level.

- Can we use  $z$  as the test statistic? Tell why or why not.
- State the null hypothesis and the alternate hypothesis.
- Show the decision rule graphically.
- Compute the value of  $z$  and state your decision regarding the null hypothesis.
- Determine and interpret the  $p$ -value.

## EXERCISES

- The following hypotheses are given.

$$H_0: \pi \leq .70$$

$$H_1: \pi > .70$$

A sample of 100 observations revealed that  $p = .75$ . At the .05 significance level, can the null hypothesis be rejected?

- State the decision rule.
  - Compute the value of the test statistic.
  - What is your decision regarding the null hypothesis?
- The following hypotheses are given.

$$H_0: \pi = .40$$

$$H_1: \pi \neq .40$$

A sample of 120 observations revealed that  $p = .30$ . At the .05 significance level, can the null hypothesis be rejected?

- State the decision rule.
- Compute the value of the test statistic.
- What is your decision regarding the null hypothesis?

*Note:* It is recommended that you use the six-step hypothesis-testing procedure in solving the following problems.

3. The U.S. Department of Transportation estimates that 10% of Americans carpool. Does that imply that 10% of cars will have two or more occupants? A sample of 300 cars traveling southbound on the New Jersey Turnpike yesterday revealed that 63 had two or more occupants. At the .01 significance level, can we conclude that 10% of cars traveling on the New Jersey Turnpike have two or more occupants?
4. A recent article reported that a job awaits only one in three new college graduates. The major reasons given were an overabundance of college graduates and a weak economy. A survey of 200 recent graduates from your school revealed that 80 students had jobs. At the .01 significance level, can we conclude that a larger proportion of students at your school have jobs?
5. Chicken Delight claims that 90% of its orders are delivered within 10 minutes of the time the order is placed. A sample of 100 orders revealed that 82 were delivered within the promised time. At the .10 significance level, can we conclude that less than 90% of the orders are delivered in less than 10 minutes?
6. Research at the University of Toledo indicates that 50% of students change their major area of study after their first year in a program. A random sample of 100 students in the College of Business revealed that 48 had changed their major area of study after their first year of the program. Has there been a significant decrease in the proportion of students who change their major after the first year in this program? Test at the .05 level of significance.

### LO15-2

Test a hypothesis about two population proportions.

## TWO-SAMPLE TESTS ABOUT PROPORTIONS

In the previous section, we considered a test of a single population proportion. However, we are often interested also in whether two sample proportions come from populations that are equal. Here are several examples.

- The vice president of human resources wishes to know whether there is a difference in the proportion of hourly employees who miss more than 5 days of work per year at the Atlanta and the Houston plants.
- General Motors is considering a new design for the Chevy Malibu. The design is shown to a group of millennials and another group of baby-boomers. General Motors wishes to know whether there is a difference in the proportion of the two groups who like the new design.
- A consultant to the airline industry is investigating the fear of flying among adults. Specifically, the consultant wishes to know whether there is a difference in the proportion of men versus women who are fearful of flying.

In the above cases, each sampled item or individual can be classified as a “success” or a “failure.” That is, in the Chevy Malibu example, each potential buyer is classified as “liking the new design” or “not liking the new design.” We then compare the proportion in the millennial group with the proportion in the baby-boomer group who indicated they liked the new design. Can we conclude that the differences are due to chance? In this study, there is no measurement obtained, only classifying the individuals or objects.

To conduct the test, we assume each sample is large enough that the normal distribution will serve as a good approximation of the binomial distribution. The test statistic follows the standard normal distribution. We compute the value of  $z$  from the following formula:

### TWO-SAMPLE TEST OF PROPORTIONS

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_c(1 - p_c)}{n_1} + \frac{p_c(1 - p_c)}{n_2}}} \quad (15-2)$$

where:

$n_1$  is the number of observations in the first sample.

$n_2$  is the number of observations in the second sample.

$p_1$  is the proportion in the first sample possessing the trait.

$p_2$  is the proportion in the second sample possessing the trait.

$p_c$  is the pooled proportion possessing the trait in the combined samples. It is called the pooled estimate of the population proportion and is computed from the following formula.

#### POOLED PROPORTION

$$p_c = \frac{x_1 + x_2}{n_1 + n_2}$$

(15-3)

where:

$x_1$  is the number possessing the trait in the first sample.

$x_2$  is the number possessing the trait in the second sample.

The following example will illustrate the two-sample test of proportions.

#### EXAMPLE

Manelli Perfume Company recently developed a new fragrance that it plans to market under the name Heavenly. A number of market studies indicate that Heavenly has very good market potential. The sales department at Manelli is particularly interested in whether there is a difference in the proportions of working and stay-at-home women who would purchase Heavenly if it were marketed.



©Digital Vision/Getty Images

#### SOLUTION

There are two independent populations, a population consisting of working women and a population consisting of stay-at-home women. Each sampled woman will be asked to smell Heavenly and indicate whether she likes the fragrance well enough to purchase a bottle.

We will use the usual six-step hypothesis-testing procedure.

**Step 1: State  $H_0$  and  $H_1$ .** In this case, the null hypothesis is: “There is no difference in the proportion of working and stay-at-home women who prefer Heavenly.” We designate  $\pi_1$  as the proportion of working women who would purchase Heavenly and  $\pi_2$  as the proportion of stay-at-home women who would purchase it. The alternate hypothesis is that the two proportions are not equal.

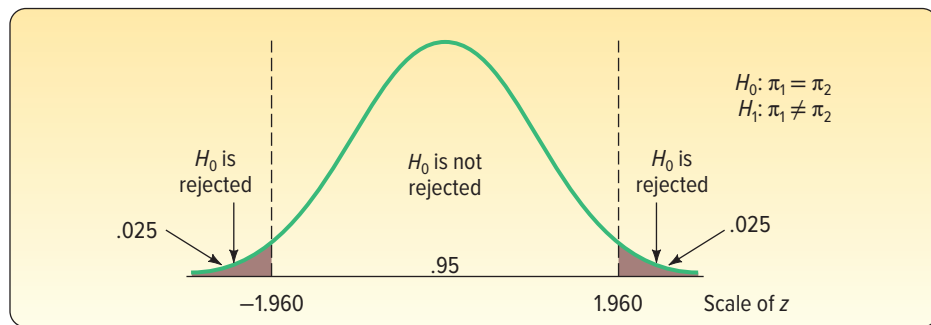
$$H_0: \pi_1 = \pi_2$$

$$H_1: \pi_1 \neq \pi_2$$

**Step 2: Select the level of significance.** We choose the .05 significance level in this example.

**Step 3: Determine the test statistic.** The two samples are sufficiently large, so we use the standard normal distribution as the test statistic. The value of the test statistic is computed using formula (15–2).

**Step 4: Formulate the decision rule.** Recall that the alternate hypothesis from Step 1 does not indicate a direction, so this is a two-tailed test. To find the critical value, go to Student’s  $t$  distribution (Appendix B.5). In the table headings, find the row labeled “Level of Significance for Two-Tailed Test” and select the column for an alpha of .05. Go to the bottom row with infinite degrees of freedom. The  $z$  critical value is 1.960, so the critical values are  $-1.960$  and  $1.960$ . As before, if the computed test statistic is less than  $-1.960$  or greater than  $1.960$ , the null hypothesis is rejected. This information is summarized in Chart 15–2.



**CHART 15–2** Decision Rules for Heavenly Fragrance Test, .05 Significance Level

**Step 5: Select a sample and make a decision.** A random sample of 100 working women revealed 19 liked the Heavenly fragrance well enough to purchase it. Similarly, a sample of 200 stay-at-home women revealed 62 liked the fragrance well enough to make a purchase. Let  $p_1$  refer to working women and  $p_2$  to stay-at-home women.

$$p_1 = \frac{x_1}{n_1} = \frac{19}{100} = .19 \quad p_2 = \frac{x_2}{n_2} = \frac{62}{200} = .31$$

The research question is whether the difference of .12 in the two sample proportions is due to chance or whether there is a difference in the proportion of working and stay-at-home women who like the Heavenly fragrance.

Next, we combine or pool the sample proportions. We use formula (15–3).

$$p_c = \frac{x_1 + x_2}{n_1 + n_2} = \frac{19 + 62}{100 + 200} = \frac{81}{300} = 0.27$$

Note that the pooled proportion is closer to .31 than to .19 because more stay-at-home women than working women were sampled.

We use formula (15–2) to find the value of the test statistic.

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_c(1 - p_c)}{n_1} + \frac{p_c(1 - p_c)}{n_2}}} = \frac{.19 - .31}{\sqrt{\frac{.27(1 - .27)}{100} + \frac{.27(1 - .27)}{200}}} = -2.207$$

The computed value of  $-2.207$  is in the area of rejection; that is, it is to the left of  $-1.960$ . Therefore, the null hypothesis is rejected at the .05 significance level. To put it another way, we reject the null hypothesis that the proportion of working women who would purchase Heavenly is equal to the proportion of stay-at-home women who would purchase Heavenly.

To find the  $p$ -value, we need to round the  $z$  test statistic from  $-2.207$  to  $-2.21$  so that we can use the table Areas under the Normal Curve in Appendix B.3. In the table, find the likelihood, or probability, of a  $z$  value less than  $-2.21$  or greater than  $2.21$ . The probability corresponding to  $2.21$  is .4864, so the likelihood of finding the value of the test statistic to be less than  $-2.21$  or greater than  $2.21$  is:

$$p\text{-value} = 2(.5000 - .4864) = 2(.0136) = .0272$$

The  $p$ -value of .0272 is less than the significance level of .05, so our decision is to reject the null hypothesis.

**Step 6: Interpret the result.** The results of the hypothesis test indicate working and stay-at-home women would purchase Heavenly at different rates or proportions.

The MegaStat add-in for Excel has a procedure to determine the value of the test statistic and compute the  $p$ -value. Notice that the MegaStat output includes the two sample proportions, the value of  $z$ , and the  $p$ -value. The difference in the  $p$ -value is rounding. The results follow.

#### Hypothesis test for two independent proportions

| $p_1$  | $p_2$  | $p_c$  |                         |
|--------|--------|--------|-------------------------|
| 0.19   | 0.31   | 0.27   | p (as decimal)          |
| 19/100 | 62/200 | 81/300 | p (as fraction)         |
| 19.    | 62.    | 81.    | X                       |
| 100    | 200    | 300    | n                       |
|        | -0.12  |        | difference              |
|        | 0.     |        | hypothesized difference |
|        | 0.0544 |        | std. error              |
|        | -2.21  |        | z                       |
|        | .0273  |        | p-value (two-tailed)    |

## SELF-REVIEW 15-2



Of 150 adults who tried a new peach-flavored Peppermint Pattie, 87 rated it excellent. Of 200 children sampled, 123 rated it excellent. Using the .10 level of significance, can we conclude that there is a significant difference in the proportion of adults and the proportion of children who rate the new flavor excellent?

- State the null hypothesis and the alternate hypothesis.
- What is the probability of a Type I error?
- Is this a one-tailed or a two-tailed test?
- What is the decision rule?
- What is the value of the test statistic?
- What is your decision regarding the null hypothesis?
- What is the  $p$ -value? Explain what it means in terms of this problem.

## EXERCISES

7. The null and alternate hypotheses are:

$$H_0: \pi_1 \leq \pi_2$$

$$H_1: \pi_1 > \pi_2$$

A sample of 100 observations from the first population indicated that  $x_1$  is 70. A sample of 150 observations from the second population revealed  $x_2$  to be 90. Use the .05 significance level to test the hypothesis.

- State the decision rule.
  - Compute the pooled proportion.
  - Compute the value of the test statistic.
  - What is your decision regarding the null hypothesis?
8. The null and alternate hypotheses are:

$$H_0: \pi_1 = \pi_2$$

$$H_1: \pi_1 \neq \pi_2$$

A sample of 200 observations from the first population indicated that  $x_1$  is 170. A sample of 150 observations from the second population revealed  $x_2$  to be 110. Use the .05 significance level to test the hypothesis.

- State the decision rule.
- Compute the pooled proportion.
- Compute the value of the test statistic.
- What is your decision regarding the null hypothesis?

*Note:* Use the six-step hypothesis-testing procedure in solving the following exercises.

9. The Damon family owns a large grape vineyard in western New York along Lake Erie. The grapevines must be sprayed at the beginning of the growing season to protect against various insects and diseases. Two new insecticides have just been marketed: Pernod 5 and Action. To test their effectiveness, three long rows were selected and sprayed with Pernod 5, and three others were sprayed with Action. When the grapes ripened, 400 of the vines treated with Pernod 5 were checked for infestation. Likewise, a sample of 400 vines sprayed with Action were checked. The results are:

| Insecticide | Number of Vines Checked (sample size) | Number of Infested Vines |
|-------------|---------------------------------------|--------------------------|
| Pernod 5    | 400                                   | 24                       |
| Action      | 400                                   | 40                       |

At the .05 significance level, can we conclude that there is a difference in the proportion of vines infested using Pernod 5 as opposed to Action?

10. GfK Research North America conducted identical surveys 5 years apart. One question asked of women was "Are most men basically kind, gentle, and thoughtful?" The earlier survey revealed that, of the 3,000 women surveyed, 2,010 said that they were. The later survey revealed 1,530 of the 3,000 women thought that men were kind, gentle, and thoughtful. At the .05 level, can we conclude that women think men are less kind, gentle, and thoughtful in the later survey compared with the earlier one?
11. A nationwide sample of influential Republicans and Democrats was asked as a part of a comprehensive survey whether they favored lowering environmental standards so that high-sulfur coal could be burned in coal-fired power plants. The results were:

|                 | Republicans | Democrats |
|-----------------|-------------|-----------|
| Number sampled  | 1,000       | 800       |
| Number in favor | 200         | 168       |

At the .02 level of significance, can we conclude that there is a larger proportion of Democrats in favor of lowering the standards? Determine the  $p$ -value.

12. The research department at the home office of New Hampshire Insurance conducts ongoing research on the causes of automobile accidents, the characteristics of the drivers, and so on. A random sample of 400 policies written on single persons revealed 120 had at least one accident in the previous three-year period. Similarly, a sample of 600 policies written on married persons revealed that 150 had been in at least one accident. At the .05 significance level, is there a significant difference in the proportions of single and married persons having an accident during a three-year period? Determine the  $p$ -value.

### LO15-3

Test a hypothesis comparing an observed set of frequencies to an expected frequency distribution.

## GOODNESS-OF-FIT TESTS: COMPARING OBSERVED AND EXPECTED FREQUENCY DISTRIBUTIONS

Next, we discuss goodness-of-fit tests that compare an observed frequency distribution to an expected frequency distribution for variables measured on a nominal or ordinal scale. For example, a life insurance company classifies its policies into four categories using a nominal variable, policy type. Policy type has four categories: whole life, level term, decreasing term, and others. The table below shows the historical relative frequency distribution of the policy types. These would be the expected frequencies.

| Policy Type     | Percent |
|-----------------|---------|
| Whole life      | 40      |
| Level term      | 25      |
| Decreasing term | 15      |
| Other           | 20      |

The insurance company wishes to compare this historical distribution with an observed distribution of policy types for a sample of 2,000 current policies. The goodness-of-fit test would determine if the current distribution of policies “fits” the historical distribution or if it has changed. A goodness-of-fit test is one of the most commonly used statistical tests.

### Hypothesis Test of Equal Expected Frequencies

Our first illustration of a goodness-of-fit test involves the case where we choose the expected frequencies to be equal. As the full name implies, the purpose of the goodness-of-fit test is to compare an observed frequency distribution to an expected frequency distribution.

#### EXAMPLE

Bubba’s Fish and Pasta is a chain of restaurants located along the Gulf Coast of Florida. Bubba, the owner, is considering adding steak to his menu. Before doing so, he decides to hire Magnolia Research LLC to conduct a survey of adults as to their favorite meal when eating out. Magnolia selected a sample 120 adults and asked each to indicate his or her favorite meal when dining out. The results are reported in Table 15–1.



**TABLE 15–1** Favorite Entrée as Selected by a Sample of 120 Adults

| Favorite Entrée | Frequency |
|-----------------|-----------|
| Chicken         | 32        |
| Fish            | 24        |
| Meat            | 35        |
| Pasta           | 29        |
| Total           | 120       |

Is it reasonable to conclude there is no preference among the four entrées?

### SOLUTION

If there is no difference in the popularity of the four entrées, we would expect the observed frequencies to be equal—or nearly equal. To put it another way, we would expect as many adults to indicate they preferred chicken as fish. Thus, any discrepancy in the observed and expected frequencies is attributed to sampling error or chance.

What is the level of measurement in this problem? Notice that when a person is selected, we can only classify the selected adult as to the entrée preferred. We do not get a reading or a measurement of any kind. The “measurement” or “classification” is based on the selected entrée. In addition, there is no natural order to the favorite entrée. No one entrée is assumed better than another. Therefore, the nominal scale is appropriate.



©EQRoy/Shutterstock

If the entrées are equally popular, we would expect 30 adults to select each meal. Why is this so? If there are 120 adults in the sample and four categories, we expect that one-fourth of those surveyed would select each entrée. So 30, found by  $120/4$ , is the expected frequency for each category, assuming there is no preference for any of the entrées. This information is summarized in Table 15–2.

An examination of the data indicates meat is the entrée selected most frequently (35 out of 120) and fish is selected least frequently (24 out of 120). Is the difference in the number of

**TABLE 15–2** Observed and Expected Frequencies for Survey of 120 Adults

| Favorite Meal | Observed Frequency, $f_o$ | Expected Frequency, $f_e$ |
|---------------|---------------------------|---------------------------|
| Chicken       | 32                        | 30                        |
| Fish          | 24                        | 30                        |
| Meat          | 35                        | 30                        |
| Pasta         | 29                        | 30                        |
| Total         | 120                       | 120                       |

times each entrée is selected due to chance, or should we conclude that the entrées are not equally preferred?

To investigate the issue, we use the six-step hypothesis-testing procedure.

**Step 1: State the null hypothesis and the alternate hypothesis.** The null hypothesis,  $H_0$ , is that there is no difference between the set of observed frequencies and the set of expected frequencies. In other words, any difference between the two sets of frequencies is attributed to sampling error. The alternate hypothesis,  $H_1$ , is that there is a difference between the observed and expected sets of frequencies. If the null hypothesis is rejected and the alternate hypothesis is accepted, we conclude the preferences are not equally distributed among the four categories.

$H_0$ : There is no difference in the proportion of adults selecting each entrée.

$H_1$ : There is a difference in the proportion of adults selecting each entrée.

**Step 2: Select the level of significance.** We selected the .05 significance level. The probability is .05 that a true null hypothesis is rejected.

**Step 3: Select the test statistic.** The test statistic follows the chi-square distribution, designated by  $\chi^2$ .

**CHI-SQUARE TEST  
STATISTIC**

$$\chi^2 = \sum \left[ \frac{(f_o - f_e)^2}{f_e} \right] \quad (15-4)$$

with  $k - 1$  degrees of freedom, where:

$k$  is the number of categories.

$f_o$  is an observed frequency in a particular category.

$f_e$  is an expected frequency in a particular category.

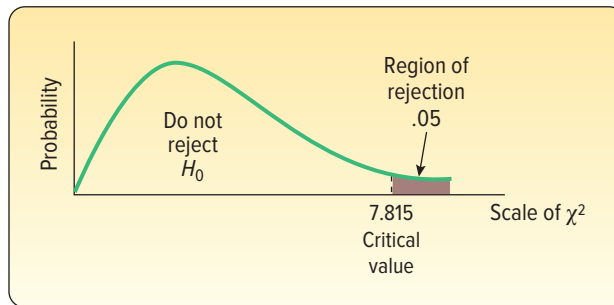
We will examine the characteristics of the chi-square distribution in more detail shortly.

**Step 4: Formulate the decision rule.** Recall that the decision rule in hypothesis testing is the value that separates the region where we do not reject  $H_0$  from the region where  $H_0$  is rejected. This number is called the *critical value*. As we will soon see, the chi-square distribution is really a family of distributions. Each distribution has a slightly different shape, depending on the number of degrees of freedom. The number of degrees of freedom is  $k - 1$ , where  $k$  is the number of categories. In this particular problem, there are four categories, the four meal entrées. Because there are four categories, there are  $k - 1 = 4 - 1 = 3$  degrees of freedom. The critical value for 3 degrees of freedom and the .05 level of significance is found in Appendix B.7. A portion of that table is shown in Table 15–3. The critical value is 7.815, found by locating 3 degrees of freedom in the left margin and then moving horizontally (to the right) and reading the critical value in the .05 column.

**TABLE 15-3** A Portion of the Chi-Square Table

| Degrees of Freedom<br><i>df</i> | Right-Tail Area |        |        |        |
|---------------------------------|-----------------|--------|--------|--------|
|                                 | .10             | .05    | .02    | .01    |
| 1                               | 2.706           | 3.841  | 5.412  | 6.635  |
| 2                               | 4.605           | 5.991  | 7.824  | 9.210  |
| 3                               | 6.251           | 7.815  | 9.837  | 11.345 |
| 4                               | 7.779           | 9.488  | 11.668 | 13.277 |
| 5                               | 9.236           | 11.070 | 13.388 | 15.086 |

The decision rule is to reject the null hypothesis if the computed value of chi-square is greater than 7.815. If it is less than or equal to 7.815, we fail to reject the null hypothesis. Chart 15-3 shows the decision rule.



**CHART 15-3** Chi-Square Probability Distribution for 3 Degrees of Freedom, Showing the Region of Rejection, .05 Level of Significance

The decision rule indicates that if there are large differences between the observed and expected frequencies, resulting in a computed  $\chi^2$  of more than 7.815, the null hypothesis should be rejected. However, if the differences between  $f_o$  and  $f_e$  are small, the computed  $\chi^2$  value will be 7.815 or less, and the null hypothesis should not be rejected. The reasoning is that such small differences between the observed and expected frequencies are probably due to chance. Remember, the 120 observations are a sample of the population.

**Step 5: Compute the value of chi-square and make a decision.** Of the 120 adults in the sample, 32 indicated their favorite entrée was chicken. The counts were reported in Table 15-1. The calculations for chi-square follow. (Note again that the expected frequencies are the same for each cell.)

Column D: Determine the differences between each  $f_o$  and  $f_e$ . That is,  $f_o - f_e$ . The sum of these differences is zero.

Column E: Square the difference between each observed and expected frequency, that is,  $(f_o - f_e)^2$ .

Column F: Divide the result for each observation by the expected frequency, that is,  $(f_o - f_e)^2 / f_e$ . Finally, sum these values. The result is the value of  $\chi^2$ , which is 2.20.

|   | A               | B     | C     | D             | E               | F                     | G              |
|---|-----------------|-------|-------|---------------|-----------------|-----------------------|----------------|
| 1 | Favorite Entrée | $f_o$ | $f_e$ | $(f_o - f_e)$ | $(f_o - f_e)^2$ | $(f_o - f_e)^2 / f_e$ |                |
| 2 | Chicken         | 32    | 30    | 2             | 4               | 0.133                 |                |
| 3 | Fish            | 24    | 30    | -6            | 36              | 1.200                 |                |
| 4 | Meat            | 35    | 30    | 5             | 25              | 0.833                 |                |
| 5 | Pasta           | 29    | 30    | -1            | 1               | 0.033                 |                |
| 6 | Total           | 120   | 120   |               |                 | 2.200                 | $\chi^2$ Value |

Source: Microsoft Excel

**STATISTICS IN ACTION**

For many years, researchers and statisticians believed that all variables were normally distributed. In fact, it was generally assumed to be a universal law. However, Karl Pearson observed that experimental data were not always normally distributed, but there was no way to prove his observations were correct. To solve this problem, Pearson discovered the chi-square statistic that basically compares an observed frequency distribution with an assumed or expected normal distribution. His discovery proved that all variables were not normally distributed.

The computed  $\chi^2$  of 2.20 is not in the rejection region. It is less than the critical value of 7.815. The decision, therefore, is to not reject the null hypothesis.

**Step 6: Interpret the results.** We conclude that the differences between the observed and the expected frequencies could be due to chance. The data do not suggest that the preferences among the four entrées are different.

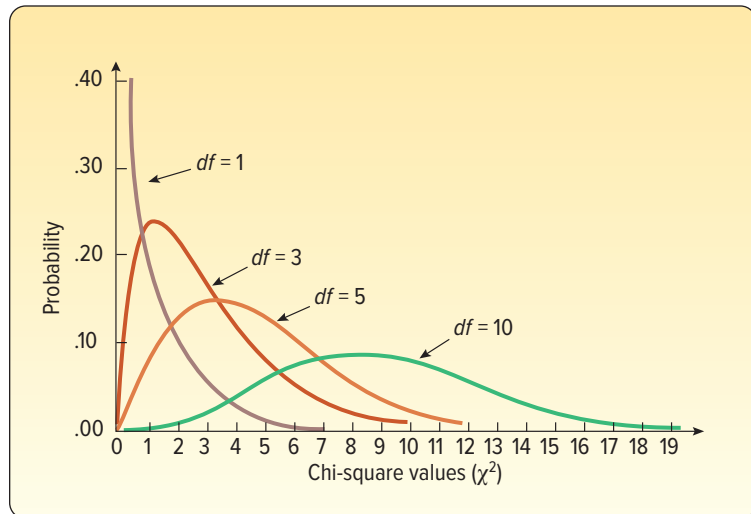
We can use MegaStat to compute the goodness-of-fit test as follows. The steps are shown in the **Software Commands** in Appendix C. The computed value of chi-square is 2.20, the same value obtained in our earlier calculations. Also note the  $p$ -value is .5319, much larger than .05.

Goodness-of-Fit Test

| observed | expected   | O - E  | (O - E) <sup>2</sup> /E | % of chisq |
|----------|------------|--------|-------------------------|------------|
| 32       | 30.000     | 2.000  | 0.133                   | 6.06       |
| 24       | 30.000     | -6.000 | 1.200                   | 54.55      |
| 35       | 30.000     | 5.000  | 0.833                   | 37.88      |
| 29       | 30.000     | -1.000 | 0.033                   | 1.52       |
| 120      | 120.000    | 0.000  | 2.200                   | 100.00     |
| 2.20     | chi-square |        |                         |            |
| 3        | df         |        |                         |            |
| .5319    | p-value    |        |                         |            |

The chi-square distribution has many applications in statistics. Its characteristics are:

- Chi-square values are never negative.** This is because the difference between  $f_o$  and  $f_e$  is squared, that is,  $(f_o - f_e)^2$ .
- There is a family of chi-square distributions.** There is a chi-square distribution for 1 degree of freedom, another for 2 degrees of freedom, another for 3 degrees of freedom, and so on. In this type of problem, the number of degrees of freedom is determined by  $k - 1$ , where  $k$  is the number of categories. Therefore, the shape of the chi-square distribution does *not* depend on the size of the sample, but on the number of categories used. For example, if 200 employees of an airline were classified into one of three categories—flight personnel, ground support, and administrative personnel—there would be  $k - 1 = 3 - 1 = 2$  degrees of freedom.
- The chi-square distribution is positively skewed.** However, as the number of degrees of freedom increases, the distribution begins to approximate the normal probability distribution. Chart 15–4 shows the distributions for selected degrees of freedom. Notice that for 10 degrees of freedom, the curve is approaching a normal distribution.



**CHART 15–4** Chi-Square Distributions for Selected Degrees of Freedom

## SELF-REVIEW 15-3



The human resources director at Georgetown Paper Inc. is concerned about absenteeism among hourly workers. She decides to sample the company records to determine whether absenteeism is distributed evenly throughout the six-day workweek. The hypotheses are:

$H_0$ : Absenteeism is evenly distributed throughout the workweek.

$H_1$ : Absenteeism is *not* evenly distributed throughout the workweek.

The sample results are:

| Number Absent |    | Number Absent |    |
|---------------|----|---------------|----|
| Monday        | 12 | Thursday      | 10 |
| Tuesday       | 9  | Friday        | 9  |
| Wednesday     | 11 | Saturday      | 9  |

- What are the numbers 12, 9, 11, 10, 9, and 9 called?
- How many categories are there?
- What is the *expected* frequency for each day?
- How many degrees of freedom are there?
- What is the chi-square critical value at the 1% significance level?
- Compute the  $\chi^2$  test statistic.
- What is the decision regarding the null hypothesis?
- Specifically, what does this indicate to the human resources director?

## EXERCISES

- In a particular chi-square goodness-of-fit test, there are four categories and 200 observations. Use the .05 significance level.
  - How many degrees of freedom are there?
  - What is the critical value of chi-square?
- In a particular chi-square goodness-of-fit test, there are six categories and 500 observations. Use the .01 significance level.
  - How many degrees of freedom are there?
  - What is the critical value of chi-square?
- The null hypothesis and the alternate hypothesis are:
 

$H_0$ : The frequencies are equal.  
 $H_1$ : The frequencies are not equal.

| Category | $f_o$ |
|----------|-------|
| A        | 10    |
| B        | 20    |
| C        | 30    |

- State the decision rule, using the .05 significance level.
  - Compute the value of chi-square.
  - What is your decision regarding  $H_0$ ?
- The null hypothesis and the alternate hypothesis are:
 

$H_0$ : The frequencies are equal.  
 $H_1$ : The frequencies are not equal.

| Category | $f_o$ |
|----------|-------|
| A        | 10    |
| B        | 20    |
| C        | 30    |
| D        | 20    |

- a. State the decision rule, using the .05 significance level.  
 b. Compute the value of chi-square.  
 c. What is your decision regarding  $H_0$ ?
17. A six-sided die is rolled 30 times and the numbers 1 through 6 appear as shown in the following frequency distribution. At the .10 significance level, can we conclude that the die is fair?

| Outcome | Frequency | Outcome | Frequency |
|---------|-----------|---------|-----------|
| 1       | 3         | 4       | 3         |
| 2       | 6         | 5       | 9         |
| 3       | 2         | 6       | 7         |

18. Classic Golf Inc. manages five courses in the Jacksonville, Florida, area. The director of golf wishes to study the number of rounds of golf played per weekday at the five courses. He gathered the following sample information. At the .05 significance level, is there a difference in the number of rounds played by day of the week?

| Day       | Rounds |
|-----------|--------|
| Monday    | 124    |
| Tuesday   | 74     |
| Wednesday | 104    |
| Thursday  | 98     |
| Friday    | 120    |

19. **FILE** A group of department store buyers viewed a new line of dresses and gave their opinions of them. The results were:

| Opinion     | Number of Buyers | Opinion     | Number of Buyers |
|-------------|------------------|-------------|------------------|
| Outstanding | 47               | Good        | 39               |
| Excellent   | 45               | Fair        | 35               |
| Very good   | 40               | Undesirable | 34               |

Because the largest number (47) indicated the new line is outstanding, the head designer thinks that this is a mandate to go into mass production of the dresses. The head sweeper (who somehow became involved in this) believes that there is not a clear mandate and claims that the opinions are evenly distributed among the six categories. He further states that the slight differences among the various counts are probably due to chance. Test the null hypothesis that there is no significant difference among the opinions of the buyers at the .01 level of significance.

20. **FILE** The safety director of a large steel mill took samples at random from company records of minor work-related accidents and classified them according to the time the accident took place.

| Time             | Number of Accidents | Time           | Number of Accidents |
|------------------|---------------------|----------------|---------------------|
| 8 up to 9 a.m.   | 6                   | 1 up to 2 p.m. | 7                   |
| 9 up to 10 a.m.  | 6                   | 2 up to 3 p.m. | 8                   |
| 10 up to 11 a.m. | 20                  | 3 up to 4 p.m. | 19                  |
| 11 up to 12 p.m. | 8                   | 4 up to 5 p.m. | 6                   |

Using the goodness-of-fit test and the .01 level of significance, determine whether the accidents are evenly distributed throughout the day. Write a brief explanation of your conclusion.

## Hypothesis Test of Unequal Expected Frequencies

The expected frequencies ( $f_e$ ) in the previous example/solution involving preferred entrées were all equal. According to the null hypothesis, it was expected that of the 120 adults in the study, an equal number would select each of the four entrées. So we expect 30 to select chicken, 30 to select fish, and so on. The chi-square test can also be used if the expected frequencies are not equal.

The following example illustrates the case of unequal frequencies and also gives a practical use of the chi-square goodness-of-fit test—namely, to find whether a local experience differs from the national experience.

### EXAMPLE

The American Hospital Administrators Association (AHAA) reports the following information concerning the number of times senior citizens are admitted to a hospital during a one-year period. Forty percent are not admitted; 30% are admitted once; 20% are admitted twice, and the remaining 10% are admitted three or more times.

A survey of 150 residents of Bartow Estates, a community devoted to active seniors located in central Florida, revealed 55 residents were not admitted during the last year, 50 were admitted to a hospital once, 32 were admitted twice, and the rest of those in the survey were admitted three or more times. Can we conclude the survey at Bartow Estates is consistent with the information reported by the AHAA? Use the .05 significance level.

### SOLUTION

We begin by organizing the above information into Table 15–4. Clearly, we cannot compare the percentages given in the AHAA study to the counts or frequencies reported for Bartow Estates residents. However, we can use the AHAA information to compute expected frequencies,  $f_e$ , for the Bartow Estates residents. According to AHAA, 40% of the seniors in their survey did not require hospitalization. Thus, if there is no difference between the national experience and the Bartow Estates study, then the expectation is that 40% of the 150 Bartow seniors surveyed, or  $f_e = 60$ , would not have been hospitalized. Further, based on the AHAA information, 30% of the 150 Bartow seniors, or  $f_e = 45$ , would be expected to be admitted once, and so on. The observed and expected frequencies for Bartow residents are given in Table 15–4.

**TABLE 15–4** Summary of Study by AHAA and a Survey of Bartow Estates

| Number of Times Admitted | AHAA Relative Frequencies | Observed Frequency of Bartow Residents ( $f_o$ ) | Expected Frequency of Bartow Residents ( $f_e$ ) |
|--------------------------|---------------------------|--|--|
| 0                        | 40%                       | 55   | $60 = (.40)(150)$                                |
| 1                        | 30%                       | 50   | $45 = (.30)(150)$                                |
| 2                        | 20%                       | 32   | $30 = (.20)(150)$                                |
| 3 or more                | 10%                       | 13   | $15 = (.10)(150)$                                |
| Total                    | 100                       | 150  |  |

The null hypothesis and the alternate hypothesis are:

$H_0$ : There is no difference between local and national experience for hospital admissions.

$H_1$ : There is a difference between local and national experience for hospital admissions.

To find the decision rule, we use Appendix B.7 and the .05 significance level. There are four admitting categories, so the degrees of freedom are  $df = 4 - 1 = 3$ . The critical value is 7.815. Therefore, the decision rule is to reject the null hypothesis if  $\chi^2 > 7.815$ . The decision rule is portrayed in Chart 15–5.

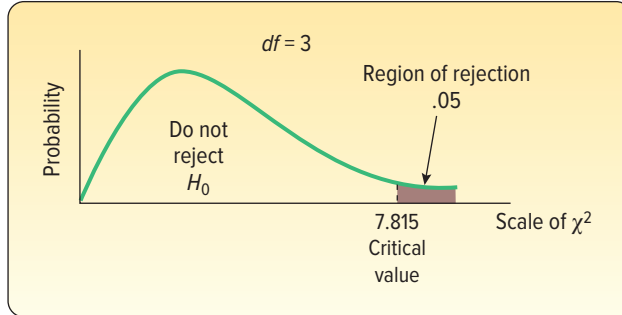


CHART 15–5 Decision Criteria for the Bartow Estates Research Study

Now to compute the chi-square test statistic:

**STATISTICS IN ACTION**

Many state governments operate lotteries to help fund education. In many lotteries, numbered balls are mixed and selected by a machine. In a Select Three game, numbered balls are selected randomly from three groups of balls numbered zero through nine. Randomness would predict that the frequency of each number is equal. How would you test if the machine ensured a random selection process? A chi-square, goodness-of-fit test could be used to investigate this question.

| Number of Times Admitted | $f_o$ | $f_e$ | $(f_o - f_e)$ | $(f_o - f_e)^2$ | $(f_o - f_e)^2 / f_e$ |
|--------------------------|-------|-------|---------------|-----------------|-----------------------|
| 0                        | 55    | 60    | -5            | 25              | 0.4167                |
| 1                        | 50    | 45    | 5             | 25              | 0.5556                |
| 2                        | 32    | 30    | 2             | 4               | 0.1333                |
| 3 or more                | 13    | 15    | -2            | 4               | 0.2667                |
| Total                    | 150   |       |               |                 | 1.3723                |

The computed value of  $\chi^2$  (1.3723) lies to the left of 7.815. Thus, we cannot reject the null hypothesis. We conclude that the survey results do not provide evidence of a difference between the local and national experience for hospital admissions.

**LO15-4**

Explain the limitations of using the chi-square statistic in goodness-of-fit tests.

**LIMITATIONS OF CHI-SQUARE**

If there is an unusually small expected frequency for a category, chi-square (if applied) might result in an erroneous conclusion. This can happen because  $f_e$  appears in the denominator, and dividing by a very small number makes the quotient quite large! Two generally accepted policies regarding small category frequencies are:

1. If there are only two cells, the *expected* frequency in each category should be at least 5. The computation of chi-square would be permissible in the following problem, involving a minimum  $f_e$  of 6.

| Individual | $f_o$ | $f_e$ |
|------------|-------|-------|
| Literate   | 641   | 642   |
| Illiterate | 7     | 6     |

2. For more than two categories, chi-square should *not* be used if more than 20% of the categories have expected frequencies less than 5. According to this policy, it would not be appropriate to use the goodness-of-fit test on the following data. Three of the seven categories, or 43%, have expected frequencies ( $f_e$ ) of less than 5.



| Level of Management      | $f_o$ | $f_e$ |
|--------------------------|-------|-------|
| Foreman                  | 30    | 32    |
| Supervisor               | 110   | 113   |
| Manager                  | 86    | 87    |
| Middle management        | 23    | 24    |
| Assistant vice president | 5     | 2     |
| Vice president           | 5     | 4     |
| Senior vice president    | 4     | 1     |
| Total                    | 263   | 263   |

To show the reason for the 20% policy, we conducted the goodness-of-fit test on the above levels-of-management data. The MegaStat output follows.

| Goodness of Fit Test |          |        |                 |            |
|----------------------|----------|--------|-----------------|------------|
| observed             | expected | O - E  | $(O - E)^2 / E$ | % of chisq |
| 30                   | 32.000   | -2.000 | 0.125           | 0.89       |
| 110                  | 113.000  | -3.000 | 0.080           | 0.57       |
| 86                   | 87.000   | -1.000 | 0.011           | 0.08       |
| 23                   | 24.000   | -1.000 | 0.042           | 0.30       |
| 5                    | 2.000    | 3.000  | 4.500           | 32.12      |
| 5                    | 4.000    | 1.000  | 0.250           | 1.78       |
| 4                    | 1.000    | 3.000  | 9.000           | 64.25      |
| 263                  | 263.000  | 0.000  | 14.008          | 100.00     |

14.01 chi-square  
6 df  
.0295 p-value

For this test at the .05 significance level,  $H_0$  is rejected if the computed value of chi-square is greater than 12.592. The computed value is 14.008, so we reject the null hypothesis that the observed and expected frequency distributions are the same. However, examine the MegaStat output critically. More than 98% of the computed chi-square value is accounted for by the three vice president categories  $[(4.500 + .250 + 9.000)/14.008 = 0.9815]$ . Logically, too much weight is being given to these categories.

The issue can be resolved by combining categories if it is logical to do so. In the above example, we combine the three vice president categories, which satisfies the 20% policy. Note that the degrees of freedom for the goodness-of-fit test change from 6 to 4.

| Level of Management | $f_o$ | $f_e$ |
|---------------------|-------|-------|
| Foreman             | 30    | 32    |
| Supervisor          | 110   | 113   |
| Manager             | 86    | 87    |
| Middle management   | 23    | 24    |
| Vice president      | 14    | 7     |
| Total               | 263   | 263   |

The computed value of chi-square with the revised categories is 7.258. See the following MegaStat output. This value is less than the critical value of 9.488 (based on 4 degrees of freedom) for the .05 significance level. The null hypothesis is, therefore, not rejected at the .05 significance level. This indicates there is not a significant difference between the observed and expected distributions.

#### Goodness-of-Fit Test

| Observed | Expected | O - E  | (O - E) <sup>2</sup> / E | % of chisq |
|----------|----------|--------|--------------------------|------------|
| 30       | 32.000   | -2.000 | 0.125                    | 1.72       |
| 110      | 113.000  | -3.000 | 0.080                    | 1.10       |
| 86       | 87.000   | -1.000 | 0.011                    | 0.16       |
| 23       | 24.000   | -1.000 | 0.042                    | 0.57       |
| 14       | 7.000    | 7.000  | 7.000                    | 96.45      |
| 263      | 263.000  | 0.000  | 7.258                    | 100.00     |

7.26 chi-square

4 df

.1229 p-value

## SELF-REVIEW 15-4



The American Accounting Association classifies accounts receivable as “current,” “late,” and “not collectible.” Industry figures show that 60% of accounts receivable are current, 30% are late, and 10% are not collectible. Massa and Barr, a law firm in Greenville, Ohio, has 500 accounts receivable; 320 are current, 120 are late, and 60 are not collectible. Are these numbers in agreement with the industry distribution? Use the .05 significance level.

## EXERCISES

21. For a particular population, a hypothesis states:

$H_0$ : Forty percent of the observations are in category A, 40% are in B, and 20% are in C.

$H_1$ : The distribution of the observations is not as described in  $H_0$ .

We took a sample of 60 observations from the population with the following results.

| Category | $f_o$ |
|----------|-------|
| A        | 30    |
| B        | 20    |
| C        | 10    |

- For the hypothesis test, state the decision rule using the .01 significance level.
  - Compute the value of chi-square.
  - What is your decision regarding  $H_0$ ?
22. The chief of security for the Mall of the Dakotas directed a study of theft. He selected a sample of 100 boxes that had been tampered with and ascertained that, for 60 of the boxes, the missing pants, shoes, and so on were attributed to shoplifting. For 30 boxes, employees had stolen the goods, and for the remaining 10 boxes, he blamed poor inventory control. In his report to the mall management, can he say

that shoplifting is *twice* as likely to be the cause of the loss as compared with either employee theft or poor inventory control and that employee theft and poor inventory control are equally likely? Use the .02 significance level.

23. From experience, the bank credit card department of Carolina Bank knows that 5% of its cardholders have had some high school, 15% have completed high school, 25% have had some college, and 55% have completed college. Of the 500 cardholders whose cards have been called in for failure to pay their charges this month, 50 had some high school, 100 had completed high school, 190 had some college, and 160 had completed college. Can we conclude that the distribution of cardholders who do not pay their charges is different from all others? Use the .01 significance level.
24. For many years, TV executives used the guideline that 30% of the audience were watching each of the traditional big three prime-time networks and 10% were watching cable stations on a weekday night. A random sample of 500 viewers in the Tampa–St. Petersburg, Florida, area last Monday night showed that 165 homes were tuned in to the ABC affiliate, 140 to the CBS affiliate, and 125 to the NBC affiliate, with the remainder viewing a cable station. At the .05 significance level, can we conclude that the guideline is still reasonable?

### LO15-5

Perform a chi-square test for independence on a contingency table.

## CONTINGENCY TABLE ANALYSIS

In Chapter 4, we discussed bivariate data, where we studied the relationship between two variables. We described a contingency table, which simultaneously summarizes two nominal-scale variables of interest. For example, a sample of students enrolled in the School of Business is classified by gender (male or female) and major (accounting, management, finance, marketing, or business analytics). This classification is based on the nominal scale because there is no natural order to the classifications.

We discussed contingency tables in Chapter 5. On page 136, we illustrated the relationship between the number of movies attended per month and the age of the attendee. We can use the chi-square distribution to test whether two nominal-scaled variables are related or not. To put it another way, is one variable *independent* of the other?

Here are some examples where we are interested in testing whether two nominal-scaled variables are related.

- Ford Motor Company operates an assembly plant in Dearborn, Michigan. The plant operates three shifts per day, 5 days a week. The quality control manager wishes to compare the quality level on the three shifts. Vehicles are classified by quality level (acceptable, unacceptable) and shift (day, afternoon, night). Is there a difference in the quality level on the three shifts? That is, is the quality of the product related to the shift when it was manufactured? Or is the quality of the product independent of the shift when it was manufactured?
- A sample of 100 drivers who were stopped for speeding violations was classified by gender and whether or not they were wearing a seat belt. For this sample, is wearing a seat belt related to gender?
- Do individuals released from federal prison make a different adjustment to civilian life if they return to their hometown or if they go elsewhere to live? The two variables are adjustment to civilian life and place of residence. Note that both variables are measured on the nominal scale.

The following example/solution provides the details of the analysis and possible conclusions.

### ▶ EXAMPLE

Rainbow Chemical Inc. employs hourly and salaried employees. The vice president of human resources surveyed 380 employees about their satisfaction level with the current health care benefits program. The employees were then

classified according to the pay type, i.e., salary or hourly. The results are shown in Table 15–5.

**TABLE 15–5** Health Care Satisfaction Level for Rainbow Chemical Employees

| Pay Type | Satisfied | Neutral | Dissatisfied | Total |
|----------|-----------|---------|--------------|-------|
| Salary   | 30        | 17      | 8            | 55    |
| Hourly   | 140       | 127     | 58           | 325   |
| Total    | 170       | 144     | 66           | 380   |

At the .05 significance level, is it reasonable to conclude that pay type and level of satisfaction with the health care benefits are related?

### SOLUTION

The first step is to state the null hypothesis and the alternate hypothesis.

$H_0$ : There is no relationship between level of satisfaction and pay type.

$H_1$ : There is a relationship between level of satisfaction and pay type.

The significance level, as requested by the HR vice president, is .05. The level of measurement for pay type is the nominal scale. The satisfaction level with health benefits is actually the ordinal scale, but we use it as a nominal-scale variable. Each sampled employee is classified by two criteria: the level of satisfaction with benefits and pay type. The information is tabulated into Table 15–5, which is called a contingency table.

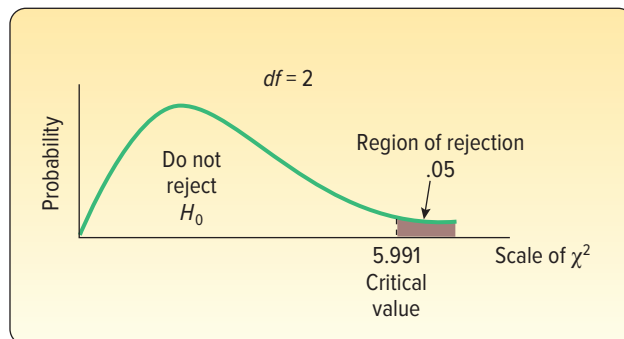
We use the chi-square distribution as the test statistic. To determine the critical value of chi-square, we calculate the degrees of freedom ( $df$ ) as:

$$df = (\text{Number of rows} - 1)(\text{Number of columns} - 1) = (r - 1)(c - 1)$$

In this example/solution, there are 2 rows and 3 columns, so there are 2 degrees of freedom.

$$df = (r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$$

To find the critical value for 2 degrees of freedom and the .05 level, refer to Appendix B.7. Move down the degrees of freedom column in the left margin to the row with 2 degrees of freedom. Move across this row to the column headed .05. At the intersection, the chi-square critical value is 5.991. The decision rule is to reject the null hypothesis if the computed value of  $\chi^2$  is greater than 5.991. See Chart 15–6.



**CHART 15–6** Chi-Square Distribution for 2 Degrees of Freedom

Next we compute the chi-square value,  $\chi^2$ , using formula (15–4). The observed frequencies,  $f_o$ , and expected frequencies,  $f_e$ , are shown in Table 15–6. How are the corresponding expected frequencies,  $f_e$ , determined? To begin, notice from Table 15–5 that 55 of the 380 Rainbow Chemical employees sampled are salaried. So the fraction of salaried employees in the sample is  $55/380 = .14474$ . *If there is no relationship* between pay type and level of satisfaction with the health care benefits program, we would expect the same fraction of the employees who are satisfied with the health care to be salaried. There are 170 employees who are satisfied with the health care program, so the expected number of satisfied employees who are salaried is 24.61, found by  $(.14474)(170)$ . Thus, the expected frequency for the upper-left cell is 24.61. Likewise, if there were no relationship between satisfaction level and pay type, we would expect .14474 of the 144 employees, or 20.84, who were neutral about the health care program to be salaried. We continue this process, filling in the remaining cells. It is not necessary to calculate each of these cell values. In fact we only need to calculate two cells. We can find the others by subtraction.

The expected frequency for any cell is determined by:

$$\text{EXPECTED FREQUENCY} \quad f_e = \frac{(\text{Row total})(\text{Column total})}{(\text{Grand total})} \quad (15-5)$$

From this formula, the expected frequency for the upper-left cell in Table 15–5 is:

$$f_e = \frac{(\text{Row total})(\text{Column total})}{(\text{Grand total})} = \frac{(55)(170)}{380} = 24.61$$

The observed frequencies,  $f_o$ , and the expected frequencies,  $f_e$ , for all of the cells in the contingency table are listed in Table 15–6. Note there are slight differences due to rounding.

**TABLE 15–6** Observed and Expected Frequencies

| Pay Type | Satisfaction Level with Health Care |        |         |        |              |       |
|----------|-------------------------------------|--------|---------|--------|--------------|-------|
|          | Satisfied                           |        | Neutral |        | Dissatisfied |       |
|          | $f_o$                               | $f_e$  | $f_o$   | $f_e$  | $f_o$        | $f_e$ |
| Salary   | 30                                  | 24.61  | 17      | 20.84  | 8            | 9.55  |
| Hourly   | 140                                 | 145.39 | 127     | 123.16 | 58           | 56.45 |
| Total    | 170                                 | 170.00 | 144     | 144.00 | 66           | 66.00 |

We use formula (15–4) to determine the value of chi-square. Starting with the upper-left cell:

$$\begin{aligned} \chi^2 &= \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(30 - 24.61)^2}{24.61} + \frac{(17 - 20.84)^2}{20.84} + \dots + \frac{(58 - 56.45)^2}{56.45} \\ &= 1.181 + .708 + \dots + .043 = 2.506 \end{aligned}$$

Because the computed value of chi-square (2.506) lies in the region to the left of 5.991, the null hypothesis is not rejected at the .05 significance level. What do we conclude? The sample data do not provide evidence that pay type and satisfaction level with health care benefits are related.

The following output is from the MegaStat Excel add-in.

Chi-Square Contingency Table Test for Independence

| Pay Type |          | Satisfaction Level with Health Care |            |              | Total  |
|----------|----------|-------------------------------------|------------|--------------|--------|
|          |          | Satisfied                           | Neutral    | Dissatisfied |        |
| Salary   | Observed | <b>30</b>                           | <b>17</b>  | <b>8</b>     | 55     |
|          | Expected | 24.61                               | 20.84      | 9.55         | 55.00  |
| Hourly   | Observed | <b>140</b>                          | <b>127</b> | <b>58</b>    | 325    |
|          | Expected | 145.39                              | 123.16     | 56.45        | 325.00 |
| Total    | Observed | 170                                 | 144        | 66           | 380    |
|          | Expected | 170.00                              | 144.00     | 66.00        | 380.00 |

2.506 chi-square  
2 df  
0.286 p-value

Observe that the value of chi-square is the same as that computed earlier, 2.506. In addition, the  $p$ -value, .286, is reported. So the probability of finding a value of the test statistic as large or larger, assuming the null hypothesis is true, is .286. The  $p$ -value also results in the same decision: do not reject the null hypothesis.

## SELF-REVIEW 15-5



A social scientist sampled 140 people and classified them according to income level and whether or not they played a state lottery in the last month. The sample information is reported below. Is it reasonable to conclude that playing the lottery is related to income level? Use the .05 significance level.

|              | Income |        |      | Total |
|--------------|--------|--------|------|-------|
|              | Low    | Middle | High |       |
| Played       | 46     | 28     | 21   | 95    |
| Did not play | 14     | 12     | 19   | 45    |
| Total        | 60     | 40     | 40   | 140   |

- What is this table called?
- State the null hypothesis and the alternate hypothesis.
- What is the decision rule?
- Determine the value of chi-square.
- Make a decision on the null hypothesis. Interpret the result.

## EXERCISES

25. **FILE** The director of advertising for the *Carolina Sun Times*, the largest newspaper in the Carolinas, is studying the relationship between the type of community in which a subscriber resides and the section of the newspaper he or she reads first. For a sample of readers, she collected the sample information in the following table.

|        | National News | Sports | Food |
|--------|---------------|--------|------|
| City   | 170           | 124    | 90   |
| Suburb | 120           | 112    | 100  |
| Rural  | 130           | 90     | 88   |

At the .05 significance level, can we conclude there is a relationship between the type of community where the person resides and the section of the paper read first?

26. **FILE** Four brands of lightbulbs are being considered for use in the final assembly area of the Ford F-150 truck plant in Dearborn, Michigan. The director of purchasing asked for samples of 100 from each manufacturer. The numbers of acceptable and unacceptable bulbs from each manufacturer are shown below. At the .05 significance level, is there a difference in the quality of the bulbs?

|              | Manufacturer |     |     |     |
|--------------|--------------|-----|-----|-----|
|              | A            | B   | C   | D   |
| Unacceptable | 12           | 8   | 5   | 11  |
| Acceptable   | 88           | 92  | 95  | 89  |
| Total        | 100          | 100 | 100 | 100 |

27. **FILE** The quality control department at Food Town Inc., a grocery chain in upstate New York, conducts a monthly check on the comparison of scanned prices to posted prices. The chart below summarizes the results of a sample of 500 items last month. Company management would like to know whether there is any relationship between error rates on regularly priced items and specially priced items. Use the .01 significance level.

|               | Regular Price | Special Price |
|---------------|---------------|---------------|
| Undercharge   | 20            | 10            |
| Overcharge    | 15            | 30            |
| Correct price | 200           | 225           |

28. **FILE** The use of cellular phones in automobiles has increased dramatically in the last few years. Of concern to traffic experts, as well as manufacturers of cellular phones, is the effect on accident rates. Is someone who is using a cellular phone more likely to be involved in a traffic accident? What is your conclusion from the following sample information? Use the .05 significance level.

|                           | Had Accident<br>in the Last Year | Did Not Have an Accident<br>in the Last Year |
|---------------------------|----------------------------------|--|
| Uses a cell phone         | 25                               | 300  |
| Does not use a cell phone | 50                               | 400  |

## CHAPTER SUMMARY

- I. This chapter considered tests of hypothesis for nominal-level data.
- II. When we sample from a single population and the variable of interest has only two possible outcomes, we call this a test of proportion.
  - A. The binomial conditions must be met.
  - B. Both  $n\pi$  and  $n(1 - \pi)$  must be at least 5.
  - C. The test statistic is

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} \quad (15-1)$$

III. We can also test whether two samples came from populations with an equal proportion of successes.

A. The two sample proportions are pooled using the following formula:

$$p_c = \frac{x_1 + x_2}{n_1 + n_2} \quad (15-3)$$

B. We compute the value of the test statistic from the following formula:

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_c(1-p_c)}{n_1} + \frac{p_c(1-p_c)}{n_2}}} \quad (15-2)$$

IV. The characteristics of the chi-square distribution are:

A. The value of chi-square is never negative.

B. The chi-square distribution is positively skewed.

C. There is a family of chi-square distributions.

1. Each time the degrees of freedom change, a new distribution is formed.

2. As the degrees of freedom increase, the distribution approaches a normal distribution.

V. A goodness-of-fit test will show whether an observed set of frequencies could have come from a hypothesized population distribution.

A. The degrees of freedom are  $k - 1$ , where  $k$  is the number of categories.

B. The formula for computing the value of chi-square is

$$\chi^2 = \sum \left[ \frac{(f_o - f_e)^2}{f_e} \right] \quad (15-4)$$

VI. A contingency table is used to test whether two traits or characteristics are related.

A. Each observation is classified according to two traits.

B. The expected frequency is determined as follows:

$$f_e = \frac{(\text{Row total})(\text{Column total})}{(\text{Grand total})} \quad (15-5)$$

C. The degrees of freedom are found by:

$$df = (\text{Rows} - 1)(\text{Columns} - 1)$$

D. The usual hypothesis testing procedure is used.

## PRONUNCIATION KEY

| SYMBOL   | MEANING              | PRONUNCIATION    |
|----------|----------------------|------------------|
| $p_c$    | Pooled proportion    | <i>p sub c</i>   |
| $\chi^2$ | Chi-square statistic | <i>ki square</i> |
| $f_o$    | Observed frequency   | <i>f sub oh</i>  |
| $f_e$    | Expected frequency   | <i>f sub e</i>   |

## CHAPTER EXERCISES

29. A coin toss is used to decide which team gets the ball first in most sports. It involves little effort and is believed to give each side the same chance. In 51 Super Bowl games, the coin toss resulted in 25 heads and 26 tails. However, the National Football Conference has correctly called the coin flip 35 times. Meanwhile, the American Football Conference has correctly called the flip only 16 times. Use the six-step hypothesis-testing procedure at the .01 significance level to test whether these data suggest that the National Football Conference has an advantage in calling the coin flip.

a. Why can you use a z-statistic as the test statistic?

b. State the null and alternate hypotheses.



- c. Make a diagram of the decision rule.
  - d. Evaluate the test statistic and make the decision.
  - e. What is the  $p$ -value and what does that imply?
30. According to a study by the American Pet Food Dealers Association, 63% of U.S. households own pets. A report is being prepared for an editorial in the *San Francisco Chronicle*. As a part of the editorial, a random sample of 300 households showed 210 own pets. Do these data disagree with the Pet Food Dealers Association's data? Use a .05 level of significance.
31. Tina Dennis is the comptroller for Meek Industries. She believes that the current cash-flow problem at Meek is due to the slow collection of accounts receivable. She believes that more than 60% of the accounts are more than 3 months in arrears. A random sample of 200 accounts showed that 140 were more than 3 months old. At the .01 significance level, can she conclude that more than 60% of the accounts are in arrears for more than three months?
32. The policy of the Suburban Transit Authority is to add a bus route if more than 55% of the potential commuters indicate they would use the particular route. A sample of 70 commuters revealed that 42 would use a proposed route from Bowman Park to the downtown area. Does the Bowman-to-downtown route meet the STA criterion? Use the .05 significance level.
33. Past experience at the Crowder Travel Agency indicated that 44% of those persons who wanted the agency to plan a vacation for them wanted to go to Europe. During the most recent season, a sampling of 1,000 persons was selected at random from the files. It was found that 480 persons wanted to go to Europe on vacation. Has there been a significant shift upward in the percentage of persons who want to go to Europe? Test at the .05 significance level.
34. Research in the gaming industry showed that 10% of all slot machines in the United States stop working each year. Short's Game Arcade has 60 slot machines and only 3 failed last year. At the .05 significance level, test whether these data contradict the research report.
- a. Why can you use a  $z$ -statistic as the test statistic?
  - b. State the null and alternate hypotheses.
  - c. Evaluate the test statistic and make the decision.
  - d. What is the  $p$ -value and what does that imply?
35. An urban planner claims that, nationally, 20% of all families renting condominiums move during a given year. A random sample of 200 families renting condominiums in the Dallas Metroplex revealed that 56 moved during the past year. At the .01 significance level, does this evidence suggest that a larger proportion of condominium owners moved in the Dallas area? Determine the  $p$ -value.
36. After a losing season, there is a great uproar to fire the head football coach. In a random sample of 200 college alumni, 80 favor keeping the coach. Test at the .05 level of significance whether the proportion of alumni who support the coach is less than 50%.
37. During the 1990s, the fatality rate for lung cancer was 80 per 100,000 people. After the turn of the century and the establishment of newer treatments and adjustment in public health advertising, a random sample of 10,000 people exhibits only six deaths due to lung cancer. Test at the .05 significance level whether that data are proof of a reduced fatality rate for lung cancer.
38. Each month the National Association of Purchasing Managers surveys purchasing managers and publishes the NAPM index. One of the questions asked on the survey is: Do you think the economy is contracting? Last month, of the 300 responding managers, 160 answered yes to the question. This month, 170 of the 290 managers indicated they felt the economy was contracting. At the .05 significance level, can we conclude that a larger proportion of the purchasing managers believe the economy is contracting this month?
39. As part of a recent survey among dual-wage-earner couples, an industrial psychologist found that 990 men out of the 1,500 surveyed believed the division of household duties was fair. A sample of 1,600 women found 970 believed the division of household duties was fair. At the .01 significance level, is it reasonable to conclude that the proportion of men who believe the division of household duties is fair is larger? What is the  $p$ -value?

40. There are two major cell phone providers in the Colorado Springs, Colorado, area, one called HTC and the other, Mountain Communications. We want to investigate the “churn rate” for each provider. Churn is the number of customers or subscribers who cut ties with a company during a given time period. At the beginning of the month, HTC had 10,000 customers; at the end of the month, HTC had 9,810 customers, for a loss of 190. For the same month, Mountain Communications started with 12,500 customers and ended the month with 12,285 customers, for a loss of 215. At the .01 significance level, is there a difference in the churn rate for the two providers?
41. The Consumer Confidence Survey is a monthly review that measures consumer confidence in the U.S. economy. It is based on a typical sample of 5,000 U.S. households. Last month 9.1% of consumers said conditions were “good.” In the prior month, only 8.5% said they were “good.” Use the six-step hypothesis-testing method at the .05 level of significance to see whether you can determine if there is an increase in the share asserting conditions are “good.” Find the  $p$ -value and explain what it means.
42. A study was conducted to determine if there was a difference in the humor content in British and American trade magazine advertisements. In an independent random sample of 270 American trade magazine advertisements, 56 were humorous. An independent random sample of 203 British trade magazines contained 52 humorous ads. Do these data provide evidence at the .05 significance level that there is a difference in the proportion of humorous ads in British versus American trade magazines?
43. The AP-Petside.com poll contacted 300 married women and 200 married men. All owned pets. One hundred of the women and 36 of the men replied that their pets are better listeners than their spouses. At the .05 significance level, is there a difference between the responses of women and men?
44. The proportion of online shoppers who actually make a purchase appears to be relatively constant over time. In 2013, among a sample of 388 online shoppers, 160 purchased merchandise. In 2017, for a sample of 307 online shoppers, 144 purchased merchandise. At the .05 level of significance, did the proportion of online shoppers change from 2013 to 2017?
45. Vehicles heading west on Front Street may turn right, turn left, or go straight ahead at Elm Street. The city traffic engineer believes that half of the vehicles will continue straight through the intersection. Of the remaining half, equal proportions will turn right and left. Two hundred vehicles were observed, with the following results. Can we conclude that the traffic engineer is correct? Use the .10 significance level.

|           | Straight | Right Turn | Left Turn |
|-----------|----------|------------|-----------|
| Frequency | 112      | 48         | 40        |

46. The publisher of a sports magazine plans to offer new subscribers one of three gifts: a sweatshirt with the logo of their favorite team, a coffee cup with the logo of their favorite team, or a pair of earrings with the logo of their favorite team. In a sample of 500 new subscribers, the number selecting each gift is reported below. At the .05 significance level, is there a preference for the gifts or should we conclude that the gifts are equally well liked?

| Gift       | Frequency |
|------------|-----------|
| Sweatshirt | 183       |
| Coffee cup | 175       |
| Earrings   | 142       |

47. In a particular metro area, there are three commercial television stations, each with its own news program from 6:00 to 6:30 p.m. According to a report in this morning’s local newspaper, a random sample of 150 viewers last night revealed 53 watched the news on WNAE (channel 5), 64 watched on WRRN (channel 11), and 33 on WSPD (channel 13). At the .05 significance level, is there a difference in the proportion of viewers watching the three channels?
48. **FILE** There are four entrances to the Government Center Building in downtown Philadelphia. The building maintenance supervisor would like to know if the entrances are equally utilized. To investigate, 400 people were observed entering the building. The number

using each entrance is reported below. At the .01 significance level, is there a difference in the use of the four entrances?

| Entrance      | Frequency |
|---------------|-----------|
| Main Street   | 140       |
| Broad Street  | 120       |
| Cherry Street | 90        |
| Walnut Street | 50        |
| Total         | 400       |

49. **FILE** The owner of a mail-order catalog would like to compare her sales with the geographic distribution of the population. According to the U.S. Bureau of the Census, 21% of the population lives in the Northeast, 24% in the Midwest, 35% in the South, and 20% in the West. Listed below is a breakdown of a sample of 400 orders randomly selected from those shipped last month. At the .01 significance level, does the distribution of the orders reflect the population?

| Region    | Frequency |
|-----------|-----------|
| Northeast | 68        |
| Midwest   | 104       |
| South     | 155       |
| West      | 73        |
| Total     | 400       |

50. **FILE** Banner Mattress and Furniture Company wishes to study the number of credit applications received per day for the last 300 days. The sample information is reported below.

| Number of Credit Applications | Frequency (Number of Days) |
|-------------------------------|----------------------------|
| 0                             | 50                         |
| 1                             | 77                         |
| 2                             | 81                         |
| 3                             | 48                         |
| 4                             | 31                         |
| 5 or more                     | 13                         |

To interpret, there were 50 days on which no credit applications were received, 77 days on which only one application was received, and so on. Would it be reasonable to conclude that the population distribution is Poisson with a mean of 2.0? Use the .05 significance level. (Hint: To find the expected frequencies, use the Poisson distribution with a mean of 2.0. Find the probability of exactly one success given a Poisson distribution with a mean of 2.0. Multiply this probability by 300 to find the expected frequency for the number of days on which there was exactly one application. Determine the expected frequency for the other days in a similar manner.)

51. **FILE** Each of the digits in a raffle is thought to have the same chance of occurrence. The table shows the frequency of each digit for consecutive drawings in a California lottery. Perform the chi-square test to see if you reject the hypothesis at the .05 significance level that the digits are from a uniform population.

| Digit | Frequency | Digit | Frequency |
|-------|-----------|-------|-----------|
| 0     | 44        | 5     | 24        |
| 1     | 32        | 6     | 31        |
| 2     | 23        | 7     | 27        |
| 3     | 27        | 8     | 28        |
| 4     | 23        | 9     | 21        |

52. **FILE** John Isaac Inc., a designer and installer of industrial signs, employs 60 people. The company recorded the type of the most recent visit to a doctor by each employee. A recent national survey found that 53% of all physician visits were to primary care physicians, 19% to medical specialists, 17% to surgical specialists, and 11% to emergency departments. Test at the .01 significance level if Isaac employees differ significantly from the survey distribution. Here are their results:

| Visit Type          | Number of Visits |
|---------------------|------------------|
| Primary care        | 29               |
| Medical specialist  | 11               |
| Surgical specialist | 16               |
| Emergency           | 4                |

53. **FILE** A survey investigated the public's attitude toward the federal deficit. Each sampled citizen was classified as to whether he or she felt the government should reduce the deficit or increase the deficit, or if the individual had no opinion. The sample results of the study by gender are reported below.

| Gender | Reduce the Deficit | Increase the Deficit | No Opinion |
|--------|--------------------|----------------------|------------|
| Female | 244                | 194                  | 68         |
| Male   | 305                | 114                  | 25         |

At the .05 significance level, is it reasonable to conclude that gender is independent of a person's position on the deficit?

54. **FILE** A study regarding the relationship between age and the amount of pressure sales personnel feel in relation to their jobs revealed the following sample information. At the .01 significance level, is there a relationship between job pressure and age?

| Age (years)  | Degree of Job Pressure |        |      |
|--------------|------------------------|--------|------|
|              | Low                    | Medium | High |
| Less than 25 | 20                     | 18     | 22   |
| 25 up to 40  | 50                     | 46     | 44   |
| 40 up to 60  | 58                     | 63     | 59   |
| 60 and older | 34                     | 43     | 43   |

55. **FILE** The claims department at Wise Insurance Company believes that younger drivers have more accidents and, therefore, should be charged higher insurance rates. Investigating a sample of 1,200 Wise policyholders revealed the following breakdown on whether a claim had been filed in the last 3 years and the age of the policyholder. Is it reasonable to conclude that there is a relationship between the age of the policyholder and whether or not the person filed a claim? Use the .05 significance level.

| Age Group   | No Claim | Claim |
|-------------|----------|-------|
| 16 up to 25 | 170      | 74    |
| 25 up to 40 | 240      | 58    |
| 40 up to 55 | 400      | 44    |
| 55 or older | 190      | 24    |
| Total       | 1,000    | 200   |

56. **FILE** A sample of employees at a large chemical plant was asked to indicate a preference for one of three pension plans. The results are given in the following table. Does it seem that there is a relationship between the pension plan selected and the job classification of the employees? Use the .01 significance level.

| Job Class  | Pension Plan |        |        |
|------------|--------------|--------|--------|
|            | Plan A       | Plan B | Plan C |
| Supervisor | 10           | 13     | 29     |
| Clerical   | 19           | 80     | 19     |
| Labor      | 81           | 57     | 22     |

57. **FILE** Did you ever purchase a bag of M&M's candies and wonder about the distribution of colors? Did you know in the beginning they were all brown? Now, peanut M&M's are 12% brown, 15% yellow, 12% red, 23% blue, 23% orange, and 15% green. A 6-oz. bag purchased at the Book Store at Coastal Carolina University had 14 brown, 13 yellow, 14 red, 12 blue, 7 orange, and 12 green. Is it reasonable to conclude that the actual distribution agrees with the expected distribution? Use the .05 significance level. Conduct your own trial. Be sure to share with your instructor.

## DATA ANALYTICS

(The data for these exercises are available in Connect.)

58. The North Valley Real Estate data report information on homes on the market.
- Determine the proportion of homes that have an attached garage. At the .05 significance level, can we conclude that more than 60% of the homes have an attached garage? What is the  $p$ -value?
  - Determine the proportion of homes that have a pool. At the .05 significance level, can we conclude that more than 60% of the homes have a pool? What is the  $p$ -value?
  - Develop a contingency table that shows whether a home has a pool and the township in which the house is located. Is there an association between the variables pool and township? Use the .05 significance level.
  - Develop a contingency table that shows whether a home has an attached garage and the township in which the home is located. Is there an association between the variables attached garage and township? Use the .05 significance level.
59. Refer to the Baseball 2016 data, which report information on the 30 Major League Baseball teams for the 2016 season. Set up a variable that divides the teams into two groups, those that had a winning season and those that did not. There are 162 games in the season, so define a winning season as having won 81 or more games. Next, find the median team salary and divide the teams into two salary groups. Let the 15 teams with the largest salaries be in one group and the 15 teams with the smallest salaries be in the other. At the .05 significance level, is there a relationship between salaries and winning?
60. Refer to the Lincolnwood School District bus data.
- Suppose we consider a bus "old" if it has been in service more than 8 years. At the .01 significance level, can we conclude that less than 40% of the district's buses are old? Report the  $p$ -value.
  - Find the median maintenance cost and the median age of the buses. Organize the data into a two-by-two contingency table, with buses above and below the median of each variable. Determine whether the age of the bus is related to the amount of the maintenance cost. Use the .05 significance level.
  - Is there a relationship between the maintenance cost and the manufacturer of the bus? Use the breakdown in part (b) for the buses above and below the median maintenance cost and the bus manufacturers to create a contingency table. Use the .05 significance level.

## PRACTICE TEST

### Part 1—Objective

1. The \_\_\_\_\_ level of measurement is required for the chi-square goodness-of-fit test.
2. To use the chi-square distribution as the test statistic, what should we assume about the population distribution? \_\_\_\_\_. (It is normally distributed; it meets the binomial conditions; or no assumption is necessary about the population distribution)
3. Which of the following is *not* a characteristic of the chi-square distribution? \_\_\_\_\_ (positively skewed, based on degrees of freedom, can have negative chi-square values)
4. In a contingency table, how many variables are summarized? \_\_\_\_\_ (two, four, fifty)
5. For a contingency table with 4 columns and 3 rows, there are \_\_\_\_\_ degrees of freedom.
6. In a contingency table, we test the null hypothesis that the variables are \_\_\_\_\_. (independent, dependent, mutually exclusive, normally distributed)
7. A sample of 100 undergraduate business students is classified by five majors. For a goodness-of-fit test, there are \_\_\_\_\_ degrees of freedom.
8. The sum of the observed and expected frequencies \_\_\_\_\_. (are the same, must be more than 30, can assume negative values, must be at least 5%)
9. In a goodness-of-fit test with 200 observations and 4 degrees of freedom, the critical value of chi-square, assuming the .05 significance level, is \_\_\_\_\_.
10. The shape of the chi-square distribution is based on the \_\_\_\_\_. (shape of the population, degrees of freedom, level of significance, level of measurement)

### Part 2—Problems

1. A recent census report indicated 65% of families have both a mother and father, 20% have only a mother, 10% have only a father, and 5% have no mother or father. A random sample of 200 children from a rural school district revealed the following:

| Mother and Father | Mother Only | Father Only | No Mother or Father | Total |
|-------------------|-------------|-------------|---------------------|-------|
| 120               | 40          | 30          | 10                  | 200   |

Is there sufficient evidence to conclude that the proportion of families with a father and/or a mother in the particular rural school district differs from the proportions reported in the recent census? Use the .05 significance level.

2. A book publisher wants to investigate the type of books selected for recreational reading by men and women. A random sample provided the following information.

| Gender | Type of Book |         |           | Total |
|--------|--------------|---------|-----------|-------|
|        | Mystery      | Romance | Self-Help |       |
| Men    | 250          | 100     | 190       | 540   |
| Women  | 130          | 170     | 200       | 500   |

At the .05 significance level, should we conclude that gender is related to the type of book selected?



# Appendixes

## APPENDIX A: DATA SETS

- A.1** Data Set 1—North Valley Real Estate Data
- A.2** Data Set 2—Baseball Statistics, 2016 Season
- A.3** Data Set 3—Lincolnville School District Bus Data
- A.4** Data Set 4—Applewood Auto Group

## APPENDIX B: TABLES

- B.1** Binomial Probability Distribution
- B.2** Poisson Distribution
- B.3** Areas under the Normal Curve
- B.4** Table of Random Numbers
- B.5** Student's  $t$  Distribution
- B.6A** Critical Values of the  $F$  Distribution, ( $\alpha = .05$ )
- B.6B** Critical Values of the  $F$  Distribution, ( $\alpha = .01$ )
- B.7** Critical Values of Chi-Square

## APPENDIX C: SOFTWARE COMMANDS

## APPENDIX D: ANSWERS TO ODD-NUMBERED CHAPTER EXERCISES & SOLUTIONS TO PRACTICE TESTS

## APPENDIX E: ANSWERS TO SELF-REVIEW



# APPENDIX A

## A.1 Data Set 1—North Valley Real Estate Data

### Variables

Record = Property identification number

Agent = Name of the real estate agent assigned to the property

Price = Market price in dollars

Size = Livable square feet of the property

Bedrooms = Number of bedrooms

Baths = Number of bathrooms

Pool = Does the home have a pool? (1 = yes, 0 = no)

Garage = Does the home have an attached garage? (1 = yes, 0 = no)

Days = Number of days of the property on the market

Township = Area where the property is located

Mortgage type = Fixed or adjustable. The fixed mortgage is a 30-year, fixed interest rate loan. The adjustable rate loan begins with an introductory interest rate of 3% for the first five years, then the interest rate is based on the current interest rates plus 1% (i.e., the interest rate AND the payment are likely to change each year after the 5th year).

Years = the number of years that the mortgage loan has been paid

FICO = the credit score of the mortgage loan holder. The highest score is 850; an average score is 680, a low score is below 680. The score reflects a person's ability to pay their debts.

Default = Is the mortgage loan in default? (1 = yes, 0 = no)

| Record | Agent    | Price  | Size | Bedrooms | Baths | Pool<br>(Yes is 1) | Garage<br>(Yes is 1) | Days | Township | Mortgage<br>type | Years | FICO | Default<br>(Yes is 1) |
|--------|----------|--------|------|----------|-------|--------------------|----------------------|------|----------|------------------|-------|------|-----------------------|
| 1      | Marty    | 206424 | 1820 | 2        | 1.5   | 1                  | 1                    | 33   | 2        | Fixed            | 2     | 824  | 0                     |
| 2      | Rose     | 346150 | 3010 | 3        | 2     | 0                  | 0                    | 36   | 4        | Fixed            | 9     | 820  | 0                     |
| 3      | Carter   | 372360 | 3210 | 4        | 3     | 0                  | 1                    | 21   | 2        | Fixed            | 18    | 819  | 0                     |
| 4      | Peterson | 310622 | 3330 | 3        | 2.5   | 1                  | 0                    | 26   | 3        | Fixed            | 17    | 817  | 0                     |
| 5      | Carter   | 496100 | 4510 | 6        | 4.5   | 0                  | 1                    | 13   | 4        | Fixed            | 17    | 816  | 0                     |
| 6      | Peterson | 294086 | 3440 | 4        | 3     | 1                  | 1                    | 31   | 4        | Fixed            | 19    | 813  | 0                     |
| 7      | Carter   | 228810 | 2630 | 4        | 2.5   | 0                  | 1                    | 39   | 4        | Adjustable       | 10    | 813  | 0                     |
| 8      | Isaacs   | 384420 | 4470 | 5        | 3.5   | 0                  | 1                    | 26   | 2        | Fixed            | 6     | 812  | 0                     |
| 9      | Peterson | 416120 | 4040 | 5        | 3.5   | 0                  | 1                    | 26   | 4        | Fixed            | 3     | 810  | 0                     |
| 10     | Isaacs   | 487494 | 4380 | 6        | 4     | 1                  | 1                    | 32   | 3        | Fixed            | 6     | 808  | 0                     |
| 11     | Rose     | 448800 | 5280 | 6        | 4     | 0                  | 1                    | 35   | 4        | Fixed            | 8     | 806  | 1                     |
| 12     | Peterson | 388960 | 4420 | 4        | 3     | 0                  | 1                    | 50   | 2        | Adjustable       | 9     | 805  | 1                     |
| 13     | Marty    | 335610 | 2970 | 3        | 2.5   | 0                  | 1                    | 25   | 3        | Adjustable       | 9     | 801  | 1                     |
| 14     | Rose     | 276000 | 2300 | 2        | 1.5   | 0                  | 0                    | 34   | 1        | Fixed            | 20    | 798  | 0                     |
| 15     | Rose     | 346421 | 2970 | 4        | 3     | 1                  | 1                    | 17   | 3        | Adjustable       | 10    | 795  | 0                     |
| 16     | Isaacs   | 453913 | 3660 | 6        | 4     | 1                  | 1                    | 12   | 3        | Fixed            | 18    | 792  | 0                     |
| 17     | Carter   | 376146 | 3290 | 5        | 3.5   | 1                  | 1                    | 28   | 2        | Adjustable       | 9     | 792  | 1                     |
| 18     | Peterson | 694430 | 5900 | 5        | 3.5   | 1                  | 1                    | 36   | 3        | Adjustable       | 10    | 788  | 0                     |
| 19     | Rose     | 251269 | 2050 | 3        | 2     | 1                  | 1                    | 38   | 3        | Fixed            | 16    | 786  | 0                     |
| 20     | Rose     | 547596 | 4920 | 6        | 4.5   | 1                  | 1                    | 37   | 5        | Fixed            | 2     | 785  | 0                     |
| 21     | Marty    | 214910 | 1950 | 2        | 1.5   | 1                  | 0                    | 20   | 4        | Fixed            | 6     | 784  | 0                     |
| 22     | Rose     | 188799 | 1950 | 2        | 1.5   | 1                  | 0                    | 52   | 1        | Fixed            | 10    | 782  | 0                     |
| 23     | Carter   | 459950 | 4680 | 4        | 3     | 1                  | 1                    | 31   | 4        | Fixed            | 8     | 781  | 0                     |
| 24     | Isaacs   | 264160 | 2540 | 3        | 2.5   | 0                  | 1                    | 40   | 1        | Fixed            | 18    | 780  | 0                     |
| 25     | Carter   | 393557 | 3180 | 4        | 3     | 1                  | 1                    | 54   | 1        | Fixed            | 20    | 776  | 0                     |
| 26     | Isaacs   | 478675 | 4660 | 5        | 3.5   | 1                  | 1                    | 26   | 5        | Adjustable       | 9     | 773  | 0                     |
| 27     | Carter   | 384020 | 4220 | 5        | 3.5   | 0                  | 1                    | 23   | 4        | Adjustable       | 9     | 772  | 1                     |
| 28     | Marty    | 313200 | 3600 | 4        | 3     | 0                  | 1                    | 31   | 3        | Fixed            | 19    | 772  | 0                     |
| 29     | Isaacs   | 274482 | 2990 | 3        | 2     | 1                  | 0                    | 37   | 3        | Fixed            | 5     | 769  | 0                     |
| 30     | Marty    | 167962 | 1920 | 2        | 1.5   | 1                  | 1                    | 31   | 5        | Fixed            | 6     | 769  | 0                     |

(continued)

## A.1 Data Set 1—North Valley Real Estate Data (continued)

| Record | Agent    | Price  | Size | Bedrooms | Baths | Pool<br>(Yes is 1) | Garage<br>(Yes is 1) | Days | Township | Mortgage<br>type | Years | FICO | Default<br>(Yes is 1) |
|--------|----------|--------|------|----------|-------|--------------------|----------------------|------|----------|------------------|-------|------|-----------------------|
| 31     | Isaacs   | 175823 | 1970 | 2        | 1.5   | 1                  | 0                    | 28   | 5        | Adjustable       | 9     | 766  | 1                     |
| 32     | Isaacs   | 226498 | 2520 | 4        | 3     | 1                  | 1                    | 28   | 3        | Fixed            | 8     | 763  | 1                     |
| 33     | Carter   | 316827 | 3150 | 4        | 3     | 1                  | 1                    | 22   | 4        | Fixed            | 2     | 759  | 1                     |
| 34     | Carter   | 189984 | 1550 | 2        | 1.5   | 1                  | 0                    | 22   | 2        | Fixed            | 17    | 758  | 0                     |
| 35     | Marty    | 366350 | 3090 | 3        | 2     | 1                  | 1                    | 23   | 3        | Fixed            | 5     | 754  | 1                     |
| 36     | Isaacs   | 416160 | 4080 | 4        | 3     | 0                  | 1                    | 25   | 4        | Fixed            | 12    | 753  | 0                     |
| 37     | Isaacs   | 308000 | 3500 | 4        | 3     | 0                  | 1                    | 37   | 2        | Fixed            | 18    | 752  | 0                     |
| 38     | Rose     | 294357 | 2620 | 4        | 3     | 1                  | 1                    | 15   | 4        | Fixed            | 10    | 751  | 0                     |
| 39     | Carter   | 337144 | 2790 | 4        | 3     | 1                  | 1                    | 19   | 3        | Fixed            | 15    | 749  | 0                     |
| 40     | Peterson | 299730 | 2910 | 3        | 2     | 0                  | 0                    | 31   | 2        | Fixed            | 13    | 748  | 0                     |
| 41     | Rose     | 445740 | 4370 | 4        | 3     | 0                  | 1                    | 19   | 3        | Fixed            | 5     | 746  | 0                     |
| 42     | Rose     | 410592 | 4200 | 4        | 3     | 1                  | 1                    | 27   | 1        | Adjustable       | 9     | 741  | 1                     |
| 43     | Peterson | 667732 | 5570 | 5        | 3.5   | 1                  | 1                    | 29   | 5        | Fixed            | 4     | 740  | 0                     |
| 44     | Rose     | 523584 | 5050 | 6        | 4     | 1                  | 1                    | 19   | 5        | Adjustable       | 10    | 739  | 0                     |
| 45     | Marty    | 336000 | 3360 | 3        | 2     | 0                  | 0                    | 32   | 3        | Fixed            | 6     | 737  | 0                     |
| 46     | Marty    | 202598 | 2270 | 3        | 2     | 1                  | 0                    | 28   | 1        | Fixed            | 10    | 737  | 0                     |
| 47     | Marty    | 326695 | 2830 | 3        | 2.5   | 1                  | 0                    | 30   | 4        | Fixed            | 8     | 736  | 0                     |
| 48     | Rose     | 321320 | 2770 | 3        | 2     | 0                  | 1                    | 23   | 4        | Fixed            | 6     | 736  | 0                     |
| 49     | Isaacs   | 246820 | 2870 | 4        | 3     | 0                  | 1                    | 27   | 5        | Fixed            | 13    | 735  | 0                     |
| 50     | Isaacs   | 546084 | 5910 | 6        | 4     | 1                  | 1                    | 35   | 5        | Adjustable       | 10    | 731  | 0                     |
| 51     | Isaacs   | 793084 | 6800 | 8        | 5.5   | 1                  | 1                    | 27   | 4        | Fixed            | 6     | 729  | 0                     |
| 52     | Isaacs   | 174528 | 1600 | 2        | 1.5   | 1                  | 0                    | 39   | 2        | Fixed            | 15    | 728  | 0                     |
| 53     | Peterson | 392554 | 3970 | 4        | 3     | 1                  | 1                    | 30   | 4        | Fixed            | 17    | 726  | 0                     |
| 54     | Peterson | 263160 | 3060 | 3        | 2     | 0                  | 1                    | 26   | 3        | Fixed            | 10    | 726  | 0                     |
| 55     | Rose     | 237120 | 1900 | 2        | 1.5   | 1                  | 0                    | 14   | 3        | Fixed            | 18    | 723  | 0                     |
| 56     | Carter   | 225750 | 2150 | 2        | 1.5   | 1                  | 1                    | 27   | 2        | Fixed            | 15    | 715  | 0                     |
| 57     | Isaacs   | 848420 | 7190 | 6        | 4     | 0                  | 1                    | 49   | 1        | Fixed            | 5     | 710  | 0                     |
| 58     | Carter   | 371956 | 3110 | 5        | 3.5   | 1                  | 1                    | 29   | 5        | Fixed            | 8     | 710  | 0                     |
| 59     | Carter   | 404538 | 3290 | 5        | 3.5   | 1                  | 1                    | 24   | 2        | Fixed            | 14    | 707  | 0                     |
| 60     | Rose     | 250090 | 2810 | 4        | 3     | 0                  | 1                    | 18   | 5        | Fixed            | 11    | 704  | 0                     |
| 61     | Peterson | 369978 | 3830 | 4        | 2.5   | 1                  | 1                    | 27   | 4        | Fixed            | 10    | 703  | 0                     |
| 62     | Peterson | 209292 | 1630 | 2        | 1.5   | 1                  | 0                    | 18   | 3        | Fixed            | 10    | 701  | 0                     |
| 63     | Isaacs   | 190032 | 1850 | 2        | 1.5   | 1                  | 1                    | 30   | 4        | Adjustable       | 2     | 675  | 0                     |
| 64     | Isaacs   | 216720 | 2520 | 3        | 2.5   | 0                  | 0                    | 2    | 4        | Adjustable       | 5     | 674  | 1                     |
| 65     | Marty    | 323417 | 3220 | 4        | 3     | 1                  | 1                    | 22   | 4        | Adjustable       | 2     | 673  | 0                     |
| 66     | Isaacs   | 316210 | 3070 | 3        | 2     | 0                  | 0                    | 30   | 1        | Adjustable       | 1     | 673  | 0                     |
| 67     | Peterson | 226054 | 2090 | 2        | 1.5   | 1                  | 1                    | 28   | 1        | Adjustable       | 6     | 670  | 0                     |
| 68     | Marty    | 183920 | 2090 | 3        | 2     | 0                  | 0                    | 30   | 2        | Adjustable       | 8     | 669  | 1                     |
| 69     | Rose     | 248400 | 2300 | 3        | 2.5   | 1                  | 1                    | 50   | 2        | Adjustable       | 4     | 667  | 0                     |
| 70     | Isaacs   | 466560 | 5760 | 5        | 3.5   | 0                  | 1                    | 42   | 4        | Adjustable       | 3     | 665  | 0                     |
| 71     | Rose     | 667212 | 6110 | 6        | 4     | 1                  | 1                    | 21   | 3        | Adjustable       | 8     | 662  | 1                     |
| 72     | Peterson | 362710 | 4370 | 4        | 2.5   | 0                  | 1                    | 24   | 1        | Adjustable       | 2     | 656  | 0                     |
| 73     | Rose     | 265440 | 3160 | 5        | 3.5   | 1                  | 1                    | 22   | 5        | Adjustable       | 3     | 653  | 0                     |
| 74     | Rose     | 706596 | 6600 | 7        | 5     | 1                  | 1                    | 40   | 3        | Adjustable       | 7     | 652  | 1                     |
| 75     | Marty    | 293700 | 3300 | 3        | 2     | 0                  | 0                    | 14   | 4        | Adjustable       | 7     | 647  | 1                     |
| 76     | Marty    | 199448 | 2330 | 2        | 1.5   | 1                  | 1                    | 25   | 3        | Adjustable       | 5     | 644  | 1                     |
| 77     | Carter   | 369533 | 4230 | 4        | 3     | 1                  | 1                    | 32   | 2        | Adjustable       | 2     | 642  | 0                     |
| 78     | Marty    | 230121 | 2030 | 2        | 1.5   | 1                  | 0                    | 21   | 2        | Adjustable       | 3     | 639  | 0                     |
| 79     | Marty    | 169000 | 1690 | 2        | 1.5   | 0                  | 0                    | 20   | 1        | Adjustable       | 7     | 639  | 1                     |
| 80     | Peterson | 190291 | 2040 | 2        | 1.5   | 1                  | 1                    | 31   | 4        | Adjustable       | 6     | 631  | 1                     |
| 81     | Rose     | 393584 | 4660 | 4        | 3     | 1                  | 1                    | 34   | 3        | Adjustable       | 7     | 630  | 1                     |
| 82     | Marty    | 363792 | 2860 | 3        | 2.5   | 1                  | 1                    | 48   | 5        | Adjustable       | 3     | 626  | 0                     |
| 83     | Carter   | 360960 | 3840 | 6        | 4.5   | 0                  | 1                    | 32   | 2        | Adjustable       | 5     | 626  | 1                     |
| 84     | Carter   | 310877 | 3180 | 3        | 2     | 1                  | 1                    | 40   | 1        | Adjustable       | 6     | 624  | 1                     |
| 85     | Peterson | 919480 | 7670 | 8        | 5.5   | 1                  | 1                    | 30   | 4        | Adjustable       | 1     | 623  | 0                     |
| 86     | Carter   | 392904 | 3400 | 3        | 2     | 1                  | 0                    | 40   | 2        | Adjustable       | 8     | 618  | 1                     |
| 87     | Carter   | 200928 | 1840 | 2        | 1.5   | 1                  | 1                    | 36   | 4        | Adjustable       | 3     | 618  | 1                     |

(continued)

## A.1 Data Set 1—North Valley Real Estate Data (concluded)

| Record | Agent    | Price  | Size | Bedrooms | Baths | Pool<br>(Yes is 1) | Garage<br>(Yes is 1) | Days | Township | Mortgage<br>type | Years | FICO | Default<br>(Yes is 1) |
|--------|----------|--------|------|----------|-------|--------------------|----------------------|------|----------|------------------|-------|------|-----------------------|
| 88     | Carter   | 537900 | 4890 | 6        | 4     | 0                  | 1                    | 23   | 1        | Adjustable       | 7     | 614  | 0                     |
| 89     | Rose     | 258120 | 2390 | 3        | 2.5   | 0                  | 1                    | 23   | 1        | Adjustable       | 6     | 614  | 1                     |
| 90     | Carter   | 558342 | 6160 | 6        | 4     | 1                  | 1                    | 24   | 3        | Adjustable       | 7     | 613  | 0                     |
| 91     | Marty    | 302720 | 3440 | 4        | 2.5   | 0                  | 1                    | 38   | 3        | Adjustable       | 3     | 609  | 1                     |
| 92     | Isaacs   | 240115 | 2220 | 2        | 1.5   | 1                  | 0                    | 39   | 5        | Adjustable       | 1     | 609  | 0                     |
| 93     | Carter   | 793656 | 6530 | 7        | 5     | 1                  | 1                    | 53   | 4        | Adjustable       | 3     | 605  | 1                     |
| 94     | Peterson | 218862 | 1930 | 2        | 1.5   | 1                  | 0                    | 58   | 4        | Adjustable       | 1     | 604  | 0                     |
| 95     | Peterson | 383081 | 3510 | 3        | 2     | 1                  | 1                    | 27   | 2        | Adjustable       | 6     | 601  | 1                     |
| 96     | Marty    | 351520 | 3380 | 3        | 2     | 0                  | 1                    | 35   | 2        | Adjustable       | 8     | 599  | 1                     |
| 97     | Peterson | 841491 | 7030 | 6        | 4     | 1                  | 1                    | 50   | 4        | Adjustable       | 8     | 596  | 1                     |
| 98     | Marty    | 336300 | 2850 | 3        | 2.5   | 0                  | 0                    | 28   | 1        | Adjustable       | 6     | 595  | 1                     |
| 99     | Isaacs   | 312863 | 3750 | 6        | 4     | 1                  | 1                    | 12   | 4        | Adjustable       | 2     | 595  | 0                     |
| 100    | Carter   | 275033 | 3060 | 3        | 2     | 1                  | 1                    | 27   | 3        | Adjustable       | 3     | 593  | 0                     |
| 101    | Peterson | 229990 | 2110 | 2        | 1.5   | 0                  | 0                    | 37   | 3        | Adjustable       | 6     | 591  | 1                     |
| 102    | Isaacs   | 195257 | 2130 | 2        | 1.5   | 1                  | 0                    | 11   | 5        | Adjustable       | 8     | 591  | 1                     |
| 103    | Marty    | 194238 | 1650 | 2        | 1.5   | 1                  | 1                    | 30   | 2        | Adjustable       | 7     | 590  | 1                     |
| 104    | Peterson | 348528 | 2740 | 4        | 3     | 1                  | 1                    | 27   | 5        | Adjustable       | 3     | 584  | 1                     |
| 105    | Peterson | 241920 | 2240 | 2        | 1.5   | 0                  | 1                    | 34   | 5        | Adjustable       | 8     | 583  | 1                     |

## A.2 Data Set 2—Baseball Statistics, 2016 Season

### Variables

- Team = Team's name
- League = American or National League
- Year Opened = First year the team's stadium was used
- Team Salary = Total team salary expressed in millions of dollars
- Attendance = Total number of people attending regular season games
- Wins = Number of regular season games won
- ERA = Team earned run average
- BA = Team batting average
- HR = Team home runs
- Year = Year of operation
- Average salary = Average annual player salary in millions of dollars

| Team          | League   | Year Opened | Team Salary | Attendance | Wins | ERA  | BA    | HR  |
|---------------|----------|-------------|-------------|------------|------|------|-------|-----|
| Arizona       | National | 1998        | 70.76       | 2036216    | 69   | 5.09 | 0.261 | 190 |
| Atlanta       | National | 1996        | 87.62       | 2020914    | 68   | 4.51 | 0.255 | 122 |
| Baltimore     | American | 1992        | 115.59      | 2172344    | 89   | 4.22 | 0.256 | 253 |
| Boston        | American | 1912        | 182.16      | 2955434    | 93   | 4.00 | 0.282 | 208 |
| Chicago Cubs  | National | 1914        | 116.65      | 3232420    | 103  | 3.15 | 0.256 | 199 |
| Chicago Sox   | American | 1991        | 98.71       | 1746293    | 78   | 4.10 | 0.257 | 168 |
| Cincinnati    | National | 2003        | 116.73      | 1894085    | 68   | 4.91 | 0.256 | 164 |
| Cleveland     | American | 1994        | 86.34       | 1591667    | 94   | 3.84 | 0.262 | 185 |
| Colorado      | National | 1995        | 98.26       | 2602524    | 75   | 4.91 | 0.275 | 204 |
| Detroit       | American | 2000        | 172.28      | 2493859    | 86   | 4.24 | 0.267 | 211 |
| Houston       | American | 2000        | 69.06       | 2306623    | 84   | 4.06 | 0.247 | 198 |
| Kansas City   | American | 1973        | 112.91      | 2557712    | 81   | 4.21 | 0.261 | 147 |
| LA Angels     | American | 1966        | 146.45      | 3016142    | 74   | 4.28 | 0.26  | 156 |
| LA Dodgers    | National | 1962        | 223.35      | 3703312    | 91   | 3.70 | 0.249 | 189 |
| Miami         | National | 2012        | 84.64       | 1712417    | 79   | 4.05 | 0.263 | 128 |
| Milwaukee     | National | 2001        | 98.68       | 2314614    | 73   | 4.08 | 0.244 | 194 |
| Minnesota     | American | 2010        | 108.26      | 1963912    | 59   | 5.08 | 0.251 | 200 |
| NY Mets       | National | 2009        | 99.63       | 2789602    | 87   | 3.58 | 0.246 | 218 |
| NY Yankees    | American | 2009        | 213.47      | 3063405    | 84   | 4.16 | 0.252 | 183 |
| Oakland       | American | 1966        | 80.28       | 1521506    | 69   | 4.51 | 0.246 | 169 |
| Philadelphia  | National | 2004        | 133.05      | 1915144    | 71   | 4.63 | 0.24  | 161 |
| Pittsburgh    | National | 2001        | 85.89       | 2249021    | 78   | 4.21 | 0.257 | 153 |
| San Diego     | National | 2004        | 126.37      | 2351426    | 68   | 4.43 | 0.235 | 177 |
| San Francisco | National | 2000        | 166.50      | 3365256    | 87   | 3.65 | 0.258 | 130 |
| Seattle       | American | 1999        | 122.71      | 2267928    | 86   | 4.00 | 0.259 | 223 |
| St. Louis     | National | 2006        | 120.30      | 3444490    | 86   | 4.08 | 0.255 | 225 |
| Tampa Bay     | American | 1990        | 73.65       | 1286163    | 68   | 4.20 | 0.243 | 216 |
| Texas         | American | 1994        | 144.31      | 2710402    | 95   | 4.37 | 0.262 | 215 |
| Toronto       | American | 1989        | 112.90      | 3392299    | 89   | 3.78 | 0.248 | 221 |
| Washington    | National | 2008        | 166.01      | 2481938    | 95   | 3.51 | 0.256 | 203 |

(continued)

## A.2 Data Set 2—Baseball Statistics, 2016 Season (*concluded*)

| Year | Average Salary (millions) |
|------|---------------------------|
| 2000 | 1.99                      |
| 2001 | 2.26                      |
| 2002 | 2.38                      |
| 2003 | 2.56                      |
| 2004 | 2.49                      |
| 2005 | 2.63                      |
| 2006 | 2.87                      |
| 2007 | 2.94                      |
| 2008 | 3.15                      |
| 2009 | 3.24                      |
| 2010 | 3.30                      |
| 2011 | 3.31                      |
| 2012 | 3.44                      |
| 2013 | 3.65                      |
| 2014 | 3.95                      |
| 2015 | 4.25                      |
| 2016 | 4.40                      |

## A.3 Data Set 3—Lincolnville School District Bus Data

### Variables

**ID** = Bus identification number

**Manufacturer** = Source of the bus (Bluebird, Keiser, or Thompson)

**Engine Type** = If the engine is diesel, then engine type = 0; if the engine is gasoline, then engine type = 1

**Capacity** = number of seats on the bus

**Maintenance Cost** = dollars spent to maintain a bus last year

**Age** = number of years since the bus left the manufacturer

**Odometer Miles** = total number of miles traveled by a bus

**Miles** = number of miles traveled since last maintenance

| ID  | Manufacturer | Engine Type<br>(0=diesel) | Capacity | Maintenance<br>Cost | Age | Odometer<br>Miles | Miles |
|-----|--------------|---------------------------|----------|---------------------|-----|-------------------|-------|
| 10  | Keiser       | 1                         | 14       | 4646                | 5   | 54375             | 11973 |
| 396 | Thompson     | 0                         | 14       | 1072                | 2   | 21858             | 11969 |
| 122 | Bluebird     | 1                         | 55       | 9394                | 10  | 116580            | 11967 |
| 751 | Keiser       | 0                         | 14       | 1078                | 2   | 22444             | 11948 |
| 279 | Bluebird     | 0                         | 55       | 1008                | 2   | 22672             | 11925 |
| 500 | Bluebird     | 1                         | 55       | 5329                | 5   | 50765             | 11922 |
| 520 | Bluebird     | 0                         | 55       | 4794                | 10  | 119130            | 11896 |
| 759 | Keiser       | 0                         | 55       | 3952                | 8   | 87872             | 11883 |
| 714 | Bluebird     | 0                         | 42       | 3742                | 7   | 73703             | 11837 |
| 875 | Bluebird     | 0                         | 55       | 4376                | 9   | 97947             | 11814 |
| 600 | Bluebird     | 0                         | 55       | 4832                | 10  | 119860            | 11800 |
| 953 | Bluebird     | 0                         | 55       | 5160                | 10  | 117700            | 11798 |
| 101 | Bluebird     | 0                         | 55       | 1955                | 4   | 41096             | 11789 |
| 358 | Bluebird     | 0                         | 55       | 2775                | 6   | 70086             | 11782 |
| 29  | Bluebird     | 1                         | 55       | 5352                | 6   | 69438             | 11781 |
| 365 | Keiser       | 0                         | 55       | 3065                | 6   | 63384             | 11778 |
| 162 | Keiser       | 1                         | 55       | 3143                | 3   | 31266             | 11758 |
| 686 | Bluebird     | 0                         | 55       | 1569                | 3   | 34674             | 11757 |
| 370 | Keiser       | 1                         | 55       | 7766                | 8   | 86528             | 11707 |
| 887 | Bluebird     | 0                         | 55       | 3743                | 8   | 93672             | 11704 |
| 464 | Bluebird     | 1                         | 55       | 2540                | 3   | 34530             | 11698 |
| 948 | Keiser       | 0                         | 42       | 4342                | 9   | 97956             | 11691 |
| 678 | Keiser       | 0                         | 55       | 3361                | 7   | 75229             | 11668 |
| 481 | Keiser       | 1                         | 6        | 3097                | 3   | 34362             | 11662 |
| 43  | Bluebird     | 1                         | 55       | 8263                | 9   | 102969            | 11615 |
| 704 | Bluebird     | 0                         | 55       | 4218                | 8   | 83424             | 11610 |
| 814 | Bluebird     | 0                         | 55       | 2028                | 4   | 40824             | 11576 |
| 39  | Bluebird     | 1                         | 55       | 5821                | 6   | 69444             | 11533 |
| 699 | Bluebird     | 1                         | 55       | 9069                | 9   | 98307             | 11518 |
| 75  | Bluebird     | 0                         | 55       | 3011                | 6   | 71970             | 11462 |
| 693 | Keiser       | 1                         | 55       | 9193                | 9   | 101889            | 11461 |
| 989 | Keiser       | 0                         | 55       | 4795                | 9   | 106605            | 11418 |
| 982 | Bluebird     | 0                         | 55       | 505                 | 1   | 10276             | 11359 |
| 321 | Bluebird     | 0                         | 42       | 2732                | 6   | 70122             | 11358 |
| 724 | Keiser       | 0                         | 42       | 3754                | 8   | 91968             | 11344 |
| 732 | Keiser       | 0                         | 42       | 4640                | 9   | 101196            | 11342 |
| 880 | Keiser       | 1                         | 55       | 8410                | 9   | 97065             | 11336 |
| 193 | Thompson     | 0                         | 14       | 5922                | 11  | 128711            | 11248 |
| 884 | Bluebird     | 0                         | 55       | 4364                | 9   | 92457             | 11231 |
| 57  | Bluebird     | 0                         | 55       | 3190                | 7   | 79240             | 11222 |
| 731 | Bluebird     | 0                         | 42       | 3213                | 6   | 68526             | 11168 |
| 61  | Keiser       | 0                         | 55       | 4139                | 9   | 103536            | 11148 |

(continued)

### A.3 Data Set 3—Lincolnville School District Bus Data (concluded)

| ID  | Manufacturer | Engine Type | Capacity | Maintenance | Age | Odometer | Miles |
|-----|--------------|-------------|----------|-------------|-----|----------|-------|
|     |              | (0=diesel)  |          | Cost        |     | Miles    |       |
| 135 | Bluebird     | 0           | 55       | 3560        | 7   | 76426    | 11127 |
| 833 | Thompson     | 0           | 14       | 3920        | 8   | 90968    | 11112 |
| 671 | Thompson     | 1           | 14       | 6733        | 8   | 89792    | 11100 |
| 692 | Bluebird     | 0           | 55       | 3770        | 8   | 93248    | 11048 |
| 200 | Bluebird     | 0           | 55       | 5168        | 10  | 103700   | 11018 |
| 754 | Keiser       | 0           | 14       | 7380        | 14  | 146860   | 11003 |
| 540 | Bluebird     | 1           | 55       | 3656        | 4   | 45284    | 10945 |
| 660 | Bluebird     | 1           | 55       | 6213        | 6   | 64434    | 10911 |
| 353 | Keiser       | 1           | 55       | 4279        | 4   | 45744    | 10902 |
| 482 | Bluebird     | 1           | 55       | 10575       | 10  | 116534   | 10802 |
| 398 | Thompson     | 0           | 6        | 4752        | 9   | 95922    | 10802 |
| 984 | Bluebird     | 0           | 55       | 3809        | 8   | 87664    | 10760 |
| 977 | Bluebird     | 0           | 55       | 3769        | 7   | 79422    | 10759 |
| 705 | Keiser       | 0           | 42       | 2152        | 4   | 47596    | 10755 |
| 767 | Keiser       | 0           | 55       | 2985        | 6   | 71538    | 10726 |
| 326 | Bluebird     | 0           | 55       | 4563        | 9   | 107343   | 10724 |
| 120 | Keiser       | 0           | 42       | 4723        | 10  | 110320   | 10674 |
| 554 | Bluebird     | 0           | 42       | 1826        | 4   | 44604    | 10662 |
| 695 | Bluebird     | 0           | 55       | 1061        | 2   | 23152    | 10633 |
| 9   | Keiser       | 1           | 55       | 3527        | 4   | 46848    | 10591 |
| 861 | Bluebird     | 1           | 55       | 9669        | 10  | 106040   | 10551 |
| 603 | Keiser       | 0           | 14       | 2116        | 4   | 44384    | 10518 |
| 156 | Thompson     | 0           | 14       | 6212        | 12  | 140460   | 10473 |
| 427 | Keiser       | 1           | 55       | 6927        | 7   | 73423    | 10355 |
| 883 | Bluebird     | 1           | 55       | 1881        | 2   | 20742    | 10344 |
| 168 | Thompson     | 1           | 14       | 7004        | 7   | 83006    | 10315 |
| 954 | Bluebird     | 0           | 42       | 5284        | 10  | 101000   | 10235 |
| 768 | Bluebird     | 0           | 42       | 3173        | 7   | 71778    | 10227 |
| 490 | Bluebird     | 1           | 55       | 10133       | 10  | 106240   | 10210 |
| 725 | Bluebird     | 0           | 55       | 2356        | 5   | 57065    | 10209 |
| 45  | Keiser       | 0           | 55       | 3124        | 6   | 60102    | 10167 |
| 38  | Keiser       | 1           | 14       | 5976        | 6   | 61662    | 10140 |
| 314 | Thompson     | 0           | 6        | 5408        | 11  | 128117   | 10128 |
| 507 | Bluebird     | 0           | 55       | 3690        | 7   | 72849    | 10095 |
| 40  | Bluebird     | 1           | 55       | 9573        | 10  | 118470   | 10081 |
| 918 | Bluebird     | 0           | 55       | 2470        | 5   | 53620    | 10075 |
| 387 | Bluebird     | 1           | 55       | 6863        | 8   | 89960    | 10055 |
| 418 | Bluebird     | 0           | 55       | 4513        | 9   | 104715   | 10000 |

## A.4 Data Set 4—Applewood Auto Group

### Variables

**Age** = the age of the buyer at the time of the purchase

**Profit** = the amount earned by the dealership on the sale of each vehicle

**Location** = the dealership where the vehicle was purchased

**Vehicle-Type** = SUV, sedan, compact, hybrid, or truck

**Previous** = the number of vehicles previously purchased at any of the four Applewood dealerships by the customer

| Age | Profit  | Location  | Vehicle-Type | Previous | Age | Profit | Location  | Vehicle-Type | Previous |
|-----|---------|-----------|--------------|----------|-----|--------|-----------|--------------|----------|
| 21  | \$1,387 | Tionesta  | Sedan        | 0        | 40  | 1,509  | Kane      | SUV          | 2        |
| 23  | 1,754   | Sheffield | SUV          | 1        | 40  | 1,638  | Sheffield | Sedan        | 0        |
| 24  | 1,817   | Sheffield | Hybrid       | 1        | 40  | 1,961  | Sheffield | Sedan        | 1        |
| 25  | 1,040   | Sheffield | Compact      | 0        | 40  | 2,127  | Olean     | Truck        | 0        |
| 26  | 1,273   | Kane      | Sedan        | 1        | 40  | 2,430  | Tionesta  | Sedan        | 1        |
| 27  | 1,529   | Sheffield | Sedan        | 1        | 41  | 1,704  | Sheffield | Sedan        | 1        |
| 27  | 3,082   | Kane      | Truck        | 0        | 41  | 1,876  | Kane      | Sedan        | 2        |
| 28  | 1,951   | Kane      | SUV          | 1        | 41  | 2,010  | Tionesta  | Sedan        | 1        |
| 28  | 2,692   | Tionesta  | Compact      | 0        | 41  | 2,165  | Tionesta  | SUV          | 0        |
| 29  | 1,206   | Sheffield | Sedan        | 0        | 41  | 2,231  | Tionesta  | SUV          | 2        |
| 29  | 1,342   | Kane      | Sedan        | 2        | 41  | 2,389  | Kane      | Truck        | 1        |
| 30  | 443     | Kane      | Sedan        | 3        | 42  | 335    | Olean     | SUV          | 1        |
| 30  | 754     | Olean     | Sedan        | 2        | 42  | 963    | Kane      | Sedan        | 0        |
| 30  | 1,621   | Sheffield | Truck        | 1        | 42  | 1,298  | Tionesta  | Sedan        | 1        |
| 31  | 870     | Tionesta  | Sedan        | 1        | 42  | 1,410  | Kane      | SUV          | 2        |
| 31  | 1,174   | Kane      | Truck        | 0        | 42  | 1,553  | Tionesta  | Compact      | 0        |
| 31  | 1,412   | Sheffield | Sedan        | 1        | 42  | 1,648  | Olean     | SUV          | 0        |
| 31  | 1,809   | Tionesta  | Sedan        | 1        | 42  | 2,071  | Kane      | SUV          | 0        |
| 31  | 2,415   | Kane      | Sedan        | 0        | 42  | 2,116  | Kane      | Compact      | 2        |
| 32  | 1,546   | Sheffield | Truck        | 3        | 43  | 1,500  | Tionesta  | Sedan        | 0        |
| 32  | 2,148   | Tionesta  | SUV          | 2        | 43  | 1,549  | Kane      | SUV          | 2        |
| 32  | 2,207   | Sheffield | Compact      | 0        | 43  | 2,348  | Tionesta  | Sedan        | 0        |
| 32  | 2,252   | Tionesta  | SUV          | 0        | 43  | 2,498  | Tionesta  | SUV          | 1        |
| 33  | 1,428   | Kane      | SUV          | 2        | 44  | 294    | Kane      | SUV          | 1        |
| 33  | 1,889   | Olean     | SUV          | 1        | 44  | 1,115  | Kane      | Truck        | 0        |
| 34  | 1,166   | Olean     | Sedan        | 1        | 44  | 1,124  | Tionesta  | Compact      | 2        |
| 34  | 1,320   | Tionesta  | Sedan        | 1        | 44  | 1,532  | Tionesta  | SUV          | 3        |
| 34  | 2,265   | Olean     | Sedan        | 0        | 44  | 1,688  | Kane      | Sedan        | 4        |
| 35  | 1,323   | Olean     | Sedan        | 2        | 44  | 1,822  | Kane      | SUV          | 0        |
| 35  | 1,761   | Kane      | Sedan        | 1        | 44  | 1,897  | Sheffield | Compact      | 0        |
| 35  | 1,919   | Tionesta  | SUV          | 1        | 44  | 2,445  | Kane      | SUV          | 0        |
| 36  | 2,357   | Kane      | SUV          | 2        | 44  | 2,886  | Olean     | SUV          | 1        |
| 36  | 2,866   | Kane      | Sedan        | 1        | 45  | 820    | Kane      | Compact      | 1        |
| 37  | 732     | Olean     | SUV          | 1        | 45  | 1,266  | Olean     | Sedan        | 0        |
| 37  | 1,464   | Olean     | Sedan        | 3        | 45  | 1,741  | Olean     | Compact      | 2        |
| 37  | 1,626   | Tionesta  | Compact      | 4        | 45  | 1,772  | Olean     | Compact      | 1        |
| 37  | 1,761   | Olean     | SUV          | 1        | 45  | 1,932  | Tionesta  | Sedan        | 1        |
| 37  | 1,915   | Tionesta  | SUV          | 2        | 45  | 2,350  | Sheffield | Compact      | 0        |
| 37  | 2,119   | Kane      | Hybrid       | 1        | 45  | 2,422  | Kane      | Sedan        | 1        |
| 38  | 1,766   | Sheffield | SUV          | 0        | 45  | 2,446  | Olean     | Compact      | 1        |
| 38  | 2,201   | Sheffield | Truck        | 2        | 46  | 369    | Olean     | Sedan        | 1        |
| 39  | 996     | Kane      | Compact      | 2        | 46  | 978    | Kane      | Sedan        | 1        |
| 39  | 2,813   | Tionesta  | SUV          | 0        | 46  | 1,238  | Sheffield | Compact      | 1        |
| 40  | 323     | Kane      | Sedan        | 0        | 46  | 1,818  | Kane      | SUV          | 0        |
| 40  | 352     | Sheffield | Compact      | 0        | 46  | 1,824  | Olean     | Truck        | 0        |
| 40  | 482     | Olean     | Sedan        | 1        | 46  | 1,907  | Olean     | Sedan        | 0        |
| 40  | 1,144   | Tionesta  | Truck        | 0        | 46  | 1,938  | Kane      | Sedan        | 0        |
| 40  | 1,485   | Sheffield | Compact      | 0        | 46  | 1,940  | Kane      | Truck        | 3        |

(continued)



## A.4 Data Set 4—Applewood Auto Group (concluded)

| Age | Profit | Location  | Vehicle-Type | Previous | Age | Profit | Location  | Vehicle-Type | Previous |
|-----|--------|-----------|--------------|----------|-----|--------|-----------|--------------|----------|
| 46  | 2,197  | Sheffield | Sedan        | 1        | 53  | 2,175  | Olean     | Sedan        | 1        |
| 46  | 2,646  | Tionesta  | Sedan        | 2        | 54  | 1,118  | Sheffield | Compact      | 1        |
| 47  | 1,461  | Kane      | Sedan        | 0        | 54  | 2,584  | Olean     | Compact      | 2        |
| 47  | 1,731  | Tionesta  | Compact      | 0        | 54  | 2,666  | Tionesta  | Truck        | 0        |
| 47  | 2,230  | Tionesta  | Sedan        | 1        | 54  | 2,991  | Tionesta  | SUV          | 0        |
| 47  | 2,341  | Sheffield | SUV          | 1        | 55  | 934    | Sheffield | Truck        | 1        |
| 47  | 3,292  | Olean     | Sedan        | 2        | 55  | 2,063  | Kane      | SUV          | 1        |
| 48  | 1,108  | Sheffield | Sedan        | 1        | 55  | 2,083  | Sheffield | Sedan        | 1        |
| 48  | 1,295  | Sheffield | SUV          | 1        | 55  | 2,856  | Olean     | Hybrid       | 1        |
| 48  | 1,344  | Sheffield | SUV          | 0        | 55  | 2,989  | Tionesta  | Compact      | 1        |
| 48  | 1,906  | Kane      | Sedan        | 1        | 56  | 910    | Sheffield | SUV          | 0        |
| 48  | 1,952  | Tionesta  | Compact      | 1        | 56  | 1,536  | Kane      | SUV          | 0        |
| 48  | 2,070  | Kane      | SUV          | 1        | 56  | 1,957  | Sheffield | SUV          | 1        |
| 48  | 2,454  | Kane      | Sedan        | 1        | 56  | 2,240  | Olean     | Sedan        | 0        |
| 49  | 1,606  | Olean     | Compact      | 0        | 56  | 2,695  | Kane      | Sedan        | 2        |
| 49  | 1,680  | Kane      | SUV          | 3        | 57  | 1,325  | Olean     | Sedan        | 1        |
| 49  | 1,827  | Tionesta  | Truck        | 3        | 57  | 2,250  | Sheffield | Sedan        | 2        |
| 49  | 1,915  | Tionesta  | SUV          | 1        | 57  | 2,279  | Sheffield | Hybrid       | 1        |
| 49  | 2,084  | Tionesta  | Sedan        | 0        | 57  | 2,626  | Sheffield | Sedan        | 2        |
| 49  | 2,639  | Sheffield | SUV          | 0        | 58  | 1,501  | Sheffield | Hybrid       | 1        |
| 50  | 842    | Kane      | SUV          | 0        | 58  | 1,752  | Kane      | Sedan        | 3        |
| 50  | 1,963  | Sheffield | Sedan        | 1        | 58  | 2,058  | Kane      | SUV          | 1        |
| 50  | 2,059  | Sheffield | Sedan        | 1        | 58  | 2,370  | Tionesta  | Compact      | 0        |
| 50  | 2,338  | Tionesta  | SUV          | 0        | 58  | 2,637  | Sheffield | SUV          | 1        |
| 50  | 3,043  | Kane      | Sedan        | 0        | 59  | 1,426  | Sheffield | Sedan        | 0        |
| 51  | 1,059  | Kane      | SUV          | 1        | 59  | 2,944  | Olean     | SUV          | 2        |
| 51  | 1,674  | Sheffield | Sedan        | 1        | 60  | 2,147  | Olean     | Compact      | 2        |
| 51  | 1,807  | Tionesta  | Sedan        | 1        | 61  | 1,973  | Kane      | SUV          | 3        |
| 51  | 2,056  | Sheffield | Hybrid       | 0        | 61  | 2,502  | Olean     | Sedan        | 0        |
| 51  | 2,236  | Tionesta  | SUV          | 2        | 62  | 783    | Sheffield | Hybrid       | 1        |
| 51  | 2,928  | Kane      | SUV          | 0        | 62  | 1,538  | Olean     | Truck        | 1        |
| 52  | 1,269  | Tionesta  | Sedan        | 1        | 63  | 2,339  | Olean     | Compact      | 1        |
| 52  | 1,717  | Sheffield | SUV          | 3        | 64  | 2,700  | Kane      | Truck        | 0        |
| 52  | 1,797  | Kane      | Sedan        | 1        | 65  | 2,222  | Kane      | Truck        | 1        |
| 52  | 1,955  | Olean     | Hybrid       | 2        | 65  | 2,597  | Sheffield | Truck        | 0        |
| 52  | 2,199  | Tionesta  | SUV          | 0        | 65  | 2,742  | Tionesta  | SUV          | 2        |
| 52  | 2,482  | Olean     | Compact      | 0        | 68  | 1,837  | Sheffield | Sedan        | 1        |
| 52  | 2,701  | Sheffield | SUV          | 0        | 69  | 2,842  | Kane      | SUV          | 0        |
| 52  | 3,210  | Olean     | Truck        | 4        | 70  | 2,434  | Olean     | Sedan        | 4        |
| 53  | 377    | Olean     | SUV          | 1        | 72  | 1,640  | Olean     | Sedan        | 1        |
| 53  | 1,220  | Olean     | Sedan        | 0        | 72  | 1,821  | Tionesta  | SUV          | 1        |
| 53  | 1,401  | Tionesta  | SUV          | 2        | 73  | 2,487  | Olean     | Compact      | 4        |

# APPENDIX B: TABLES

## B.1 Binomial Probability Distribution

| $n = 1$<br>Probability |       |       |       |       |       |       |       |       |       |       |       |
|------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $x$                    | 0.05  | 0.10  | 0.20  | 0.30  | 0.40  | 0.50  | 0.60  | 0.70  | 0.80  | 0.90  | 0.95  |
| 0                      | 0.950 | 0.900 | 0.800 | 0.700 | 0.600 | 0.500 | 0.400 | 0.300 | 0.200 | 0.100 | 0.050 |
| 1                      | 0.050 | 0.100 | 0.200 | 0.300 | 0.400 | 0.500 | 0.600 | 0.700 | 0.800 | 0.900 | 0.950 |

| $n = 2$<br>Probability |       |       |       |       |       |       |       |       |       |       |       |
|------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $x$                    | 0.05  | 0.10  | 0.20  | 0.30  | 0.40  | 0.50  | 0.60  | 0.70  | 0.80  | 0.90  | 0.95  |
| 0                      | 0.903 | 0.810 | 0.640 | 0.490 | 0.360 | 0.250 | 0.160 | 0.090 | 0.040 | 0.010 | 0.003 |
| 1                      | 0.095 | 0.180 | 0.320 | 0.420 | 0.480 | 0.500 | 0.480 | 0.420 | 0.320 | 0.180 | 0.095 |
| 2                      | 0.003 | 0.010 | 0.040 | 0.090 | 0.160 | 0.250 | 0.360 | 0.490 | 0.640 | 0.810 | 0.903 |

| $n = 3$<br>Probability |       |       |       |       |       |       |       |       |       |       |       |
|------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $x$                    | 0.05  | 0.10  | 0.20  | 0.30  | 0.40  | 0.50  | 0.60  | 0.70  | 0.80  | 0.90  | 0.95  |
| 0                      | 0.857 | 0.729 | 0.512 | 0.343 | 0.216 | 0.125 | 0.064 | 0.027 | 0.008 | 0.001 | 0.000 |
| 1                      | 0.135 | 0.243 | 0.384 | 0.441 | 0.432 | 0.375 | 0.288 | 0.189 | 0.096 | 0.027 | 0.007 |
| 2                      | 0.007 | 0.027 | 0.096 | 0.189 | 0.288 | 0.375 | 0.432 | 0.441 | 0.384 | 0.243 | 0.135 |
| 3                      | 0.000 | 0.001 | 0.008 | 0.027 | 0.064 | 0.125 | 0.216 | 0.343 | 0.512 | 0.729 | 0.857 |

| $n = 4$<br>Probability |       |       |       |       |       |       |       |       |       |       |       |
|------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $x$                    | 0.05  | 0.10  | 0.20  | 0.30  | 0.40  | 0.50  | 0.60  | 0.70  | 0.80  | 0.90  | 0.95  |
| 0                      | 0.815 | 0.656 | 0.410 | 0.240 | 0.130 | 0.063 | 0.026 | 0.008 | 0.002 | 0.000 | 0.000 |
| 1                      | 0.171 | 0.292 | 0.410 | 0.412 | 0.346 | 0.250 | 0.154 | 0.076 | 0.026 | 0.004 | 0.000 |
| 2                      | 0.014 | 0.049 | 0.154 | 0.265 | 0.346 | 0.375 | 0.346 | 0.265 | 0.154 | 0.049 | 0.014 |
| 3                      | 0.000 | 0.004 | 0.026 | 0.076 | 0.154 | 0.250 | 0.346 | 0.412 | 0.410 | 0.292 | 0.171 |
| 4                      | 0.000 | 0.000 | 0.002 | 0.008 | 0.026 | 0.063 | 0.130 | 0.240 | 0.410 | 0.656 | 0.815 |

| $n = 5$<br>Probability |       |       |       |       |       |       |       |       |       |       |       |
|------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $x$                    | 0.05  | 0.10  | 0.20  | 0.30  | 0.40  | 0.50  | 0.60  | 0.70  | 0.80  | 0.90  | 0.95  |
| 0                      | 0.774 | 0.590 | 0.328 | 0.168 | 0.078 | 0.031 | 0.010 | 0.002 | 0.000 | 0.000 | 0.000 |
| 1                      | 0.204 | 0.328 | 0.410 | 0.360 | 0.259 | 0.156 | 0.077 | 0.028 | 0.006 | 0.000 | 0.000 |
| 2                      | 0.021 | 0.073 | 0.205 | 0.309 | 0.346 | 0.313 | 0.230 | 0.132 | 0.051 | 0.008 | 0.001 |
| 3                      | 0.001 | 0.008 | 0.051 | 0.132 | 0.230 | 0.313 | 0.346 | 0.309 | 0.205 | 0.073 | 0.021 |
| 4                      | 0.000 | 0.000 | 0.006 | 0.028 | 0.077 | 0.156 | 0.259 | 0.360 | 0.410 | 0.328 | 0.204 |
| 5                      | 0.000 | 0.000 | 0.000 | 0.002 | 0.010 | 0.031 | 0.078 | 0.168 | 0.328 | 0.590 | 0.774 |

(continued)

## B.1 Binomial Probability Distribution (*continued*)

$n = 6$   
Probability

| $x$ | 0.05  | 0.10  | 0.20  | 0.30  | 0.40  | 0.50  | 0.60  | 0.70  | 0.80  | 0.90  | 0.95  |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0   | 0.735 | 0.531 | 0.262 | 0.118 | 0.047 | 0.016 | 0.004 | 0.001 | 0.000 | 0.000 | 0.000 |
| 1   | 0.232 | 0.354 | 0.393 | 0.303 | 0.187 | 0.094 | 0.037 | 0.010 | 0.002 | 0.000 | 0.000 |
| 2   | 0.031 | 0.098 | 0.246 | 0.324 | 0.311 | 0.234 | 0.138 | 0.060 | 0.015 | 0.001 | 0.000 |
| 3   | 0.002 | 0.015 | 0.082 | 0.185 | 0.276 | 0.313 | 0.276 | 0.185 | 0.082 | 0.015 | 0.002 |
| 4   | 0.000 | 0.001 | 0.015 | 0.060 | 0.138 | 0.234 | 0.311 | 0.324 | 0.246 | 0.098 | 0.031 |
| 5   | 0.000 | 0.000 | 0.002 | 0.010 | 0.037 | 0.094 | 0.187 | 0.303 | 0.393 | 0.531 | 0.735 |
| 6   | 0.000 | 0.000 | 0.000 | 0.001 | 0.004 | 0.016 | 0.047 | 0.118 | 0.262 | 0.531 | 0.735 |

$n = 7$   
Probability

| $x$ | 0.05  | 0.10  | 0.20  | 0.30  | 0.40  | 0.50  | 0.60  | 0.70  | 0.80  | 0.90  | 0.95  |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0   | 0.698 | 0.478 | 0.210 | 0.082 | 0.028 | 0.008 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 |
| 1   | 0.257 | 0.372 | 0.367 | 0.247 | 0.131 | 0.055 | 0.017 | 0.004 | 0.000 | 0.000 | 0.000 |
| 2   | 0.041 | 0.124 | 0.275 | 0.318 | 0.261 | 0.164 | 0.077 | 0.025 | 0.004 | 0.000 | 0.000 |
| 3   | 0.004 | 0.023 | 0.115 | 0.227 | 0.290 | 0.273 | 0.194 | 0.097 | 0.029 | 0.003 | 0.000 |
| 4   | 0.000 | 0.003 | 0.029 | 0.097 | 0.194 | 0.273 | 0.290 | 0.227 | 0.115 | 0.023 | 0.004 |
| 5   | 0.000 | 0.000 | 0.004 | 0.025 | 0.077 | 0.164 | 0.261 | 0.318 | 0.275 | 0.124 | 0.041 |
| 6   | 0.000 | 0.000 | 0.000 | 0.004 | 0.017 | 0.055 | 0.131 | 0.247 | 0.367 | 0.372 | 0.257 |
| 7   | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.008 | 0.028 | 0.082 | 0.210 | 0.478 | 0.698 |

$n = 8$   
Probability

| $x$ | 0.05  | 0.10  | 0.20  | 0.30  | 0.40  | 0.50  | 0.60  | 0.70  | 0.80  | 0.90  | 0.95  |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0   | 0.663 | 0.430 | 0.168 | 0.058 | 0.017 | 0.004 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 |
| 1   | 0.279 | 0.383 | 0.336 | 0.198 | 0.090 | 0.031 | 0.008 | 0.001 | 0.000 | 0.000 | 0.000 |
| 2   | 0.051 | 0.149 | 0.294 | 0.296 | 0.209 | 0.109 | 0.041 | 0.010 | 0.001 | 0.000 | 0.000 |
| 3   | 0.005 | 0.033 | 0.147 | 0.254 | 0.279 | 0.219 | 0.124 | 0.047 | 0.009 | 0.000 | 0.000 |
| 4   | 0.000 | 0.005 | 0.046 | 0.136 | 0.232 | 0.273 | 0.232 | 0.136 | 0.046 | 0.005 | 0.000 |
| 5   | 0.000 | 0.000 | 0.009 | 0.047 | 0.124 | 0.219 | 0.279 | 0.254 | 0.147 | 0.033 | 0.005 |
| 6   | 0.000 | 0.000 | 0.001 | 0.010 | 0.041 | 0.109 | 0.209 | 0.296 | 0.294 | 0.149 | 0.051 |
| 7   | 0.000 | 0.000 | 0.000 | 0.001 | 0.008 | 0.031 | 0.090 | 0.198 | 0.336 | 0.383 | 0.279 |
| 8   | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.004 | 0.017 | 0.058 | 0.168 | 0.430 | 0.663 |

(*continued*)

## B.1 Binomial Probability Distribution (*continued*)

| $n = 9$<br>Probability |       |       |       |       |       |       |       |       |       |       |       |
|------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $x$                    | 0.05  | 0.10  | 0.20  | 0.30  | 0.40  | 0.50  | 0.60  | 0.70  | 0.80  | 0.90  | 0.95  |
| 0                      | 0.630 | 0.387 | 0.134 | 0.040 | 0.010 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 1                      | 0.299 | 0.387 | 0.302 | 0.156 | 0.060 | 0.018 | 0.004 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2                      | 0.063 | 0.172 | 0.302 | 0.267 | 0.161 | 0.070 | 0.021 | 0.004 | 0.000 | 0.000 | 0.000 |
| 3                      | 0.008 | 0.045 | 0.176 | 0.267 | 0.251 | 0.164 | 0.074 | 0.021 | 0.003 | 0.000 | 0.000 |
| 4                      | 0.001 | 0.007 | 0.066 | 0.172 | 0.251 | 0.246 | 0.167 | 0.074 | 0.017 | 0.001 | 0.000 |
| 5                      | 0.000 | 0.001 | 0.017 | 0.074 | 0.167 | 0.246 | 0.251 | 0.172 | 0.066 | 0.007 | 0.001 |
| 6                      | 0.000 | 0.000 | 0.003 | 0.021 | 0.074 | 0.164 | 0.251 | 0.267 | 0.176 | 0.045 | 0.008 |
| 7                      | 0.000 | 0.000 | 0.000 | 0.004 | 0.021 | 0.070 | 0.161 | 0.267 | 0.302 | 0.172 | 0.063 |
| 8                      | 0.000 | 0.000 | 0.000 | 0.000 | 0.004 | 0.018 | 0.060 | 0.156 | 0.302 | 0.387 | 0.299 |
| 9                      | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.010 | 0.040 | 0.134 | 0.387 | 0.630 |

| $n = 10$<br>Probability |       |       |       |       |       |       |       |       |       |       |       |
|-------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $x$                     | 0.05  | 0.10  | 0.20  | 0.30  | 0.40  | 0.50  | 0.60  | 0.70  | 0.80  | 0.90  | 0.95  |
| 0                       | 0.599 | 0.349 | 0.107 | 0.028 | 0.006 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 1                       | 0.315 | 0.387 | 0.268 | 0.121 | 0.040 | 0.010 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2                       | 0.075 | 0.194 | 0.302 | 0.233 | 0.121 | 0.044 | 0.011 | 0.001 | 0.000 | 0.000 | 0.000 |
| 3                       | 0.010 | 0.057 | 0.201 | 0.267 | 0.215 | 0.117 | 0.042 | 0.009 | 0.001 | 0.000 | 0.000 |
| 4                       | 0.001 | 0.011 | 0.088 | 0.200 | 0.251 | 0.205 | 0.111 | 0.037 | 0.006 | 0.000 | 0.000 |
| 5                       | 0.000 | 0.001 | 0.026 | 0.103 | 0.201 | 0.246 | 0.201 | 0.103 | 0.026 | 0.001 | 0.000 |
| 6                       | 0.000 | 0.000 | 0.006 | 0.037 | 0.111 | 0.205 | 0.251 | 0.200 | 0.088 | 0.011 | 0.001 |
| 7                       | 0.000 | 0.000 | 0.001 | 0.009 | 0.042 | 0.117 | 0.215 | 0.267 | 0.201 | 0.057 | 0.010 |
| 8                       | 0.000 | 0.000 | 0.000 | 0.001 | 0.011 | 0.044 | 0.121 | 0.233 | 0.302 | 0.194 | 0.075 |
| 9                       | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.010 | 0.040 | 0.121 | 0.268 | 0.387 | 0.315 |
| 10                      | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.006 | 0.028 | 0.107 | 0.349 | 0.599 |

| $n = 11$<br>Probability |       |       |       |       |       |       |       |       |       |       |       |
|-------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $x$                     | 0.05  | 0.10  | 0.20  | 0.30  | 0.40  | 0.50  | 0.60  | 0.70  | 0.80  | 0.90  | 0.95  |
| 0                       | 0.569 | 0.314 | 0.086 | 0.020 | 0.004 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 1                       | 0.329 | 0.384 | 0.236 | 0.093 | 0.027 | 0.005 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2                       | 0.087 | 0.213 | 0.295 | 0.200 | 0.089 | 0.027 | 0.005 | 0.001 | 0.000 | 0.000 | 0.000 |
| 3                       | 0.014 | 0.071 | 0.221 | 0.257 | 0.177 | 0.081 | 0.023 | 0.004 | 0.000 | 0.000 | 0.000 |
| 4                       | 0.001 | 0.016 | 0.111 | 0.220 | 0.236 | 0.161 | 0.070 | 0.017 | 0.002 | 0.000 | 0.000 |
| 5                       | 0.000 | 0.002 | 0.039 | 0.132 | 0.221 | 0.226 | 0.147 | 0.057 | 0.010 | 0.000 | 0.000 |
| 6                       | 0.000 | 0.000 | 0.010 | 0.057 | 0.147 | 0.226 | 0.221 | 0.132 | 0.039 | 0.002 | 0.000 |
| 7                       | 0.000 | 0.000 | 0.002 | 0.017 | 0.070 | 0.161 | 0.236 | 0.220 | 0.111 | 0.016 | 0.001 |
| 8                       | 0.000 | 0.000 | 0.000 | 0.004 | 0.023 | 0.081 | 0.177 | 0.257 | 0.221 | 0.071 | 0.014 |
| 9                       | 0.000 | 0.000 | 0.000 | 0.001 | 0.005 | 0.027 | 0.089 | 0.200 | 0.295 | 0.213 | 0.087 |
| 10                      | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.005 | 0.027 | 0.093 | 0.236 | 0.384 | 0.329 |
| 11                      | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.004 | 0.020 | 0.086 | 0.314 | 0.569 |

(*continued*)

## B.1 Binomial Probability Distribution (*continued*)

| $n = 12$<br>Probability |       |       |       |       |       |       |       |       |       |       |       |
|-------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $x$                     | 0.05  | 0.10  | 0.20  | 0.30  | 0.40  | 0.50  | 0.60  | 0.70  | 0.80  | 0.90  | 0.95  |
| 0                       | 0.540 | 0.282 | 0.069 | 0.014 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 1                       | 0.341 | 0.377 | 0.206 | 0.071 | 0.017 | 0.003 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2                       | 0.099 | 0.230 | 0.283 | 0.168 | 0.064 | 0.016 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 |
| 3                       | 0.017 | 0.085 | 0.236 | 0.240 | 0.142 | 0.054 | 0.012 | 0.001 | 0.000 | 0.000 | 0.000 |
| 4                       | 0.002 | 0.021 | 0.133 | 0.231 | 0.213 | 0.121 | 0.042 | 0.008 | 0.001 | 0.000 | 0.000 |
| 5                       | 0.000 | 0.004 | 0.053 | 0.158 | 0.227 | 0.193 | 0.101 | 0.029 | 0.003 | 0.000 | 0.000 |
| 6                       | 0.000 | 0.000 | 0.016 | 0.079 | 0.177 | 0.226 | 0.177 | 0.079 | 0.016 | 0.000 | 0.000 |
| 7                       | 0.000 | 0.000 | 0.003 | 0.029 | 0.101 | 0.193 | 0.227 | 0.158 | 0.053 | 0.004 | 0.000 |
| 8                       | 0.000 | 0.000 | 0.001 | 0.008 | 0.042 | 0.121 | 0.213 | 0.231 | 0.133 | 0.021 | 0.002 |
| 9                       | 0.000 | 0.000 | 0.000 | 0.001 | 0.012 | 0.054 | 0.142 | 0.240 | 0.236 | 0.085 | 0.017 |
| 10                      | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.016 | 0.064 | 0.168 | 0.283 | 0.230 | 0.099 |
| 11                      | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.003 | 0.017 | 0.071 | 0.206 | 0.377 | 0.341 |
| 12                      | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.014 | 0.069 | 0.282 | 0.540 |

| $n = 13$<br>Probability |       |       |       |       |       |       |       |       |       |       |       |
|-------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $x$                     | 0.05  | 0.10  | 0.20  | 0.30  | 0.40  | 0.50  | 0.60  | 0.70  | 0.80  | 0.90  | 0.95  |
| 0                       | 0.513 | 0.254 | 0.055 | 0.010 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 1                       | 0.351 | 0.367 | 0.179 | 0.054 | 0.011 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2                       | 0.111 | 0.245 | 0.268 | 0.139 | 0.045 | 0.010 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 |
| 3                       | 0.021 | 0.100 | 0.246 | 0.218 | 0.111 | 0.035 | 0.006 | 0.001 | 0.000 | 0.000 | 0.000 |
| 4                       | 0.003 | 0.028 | 0.154 | 0.234 | 0.184 | 0.087 | 0.024 | 0.003 | 0.000 | 0.000 | 0.000 |
| 5                       | 0.000 | 0.006 | 0.069 | 0.180 | 0.221 | 0.157 | 0.066 | 0.014 | 0.001 | 0.000 | 0.000 |
| 6                       | 0.000 | 0.001 | 0.023 | 0.103 | 0.197 | 0.209 | 0.131 | 0.044 | 0.006 | 0.000 | 0.000 |
| 7                       | 0.000 | 0.000 | 0.006 | 0.044 | 0.131 | 0.209 | 0.197 | 0.103 | 0.023 | 0.001 | 0.000 |
| 8                       | 0.000 | 0.000 | 0.001 | 0.014 | 0.066 | 0.157 | 0.221 | 0.180 | 0.069 | 0.006 | 0.000 |
| 9                       | 0.000 | 0.000 | 0.000 | 0.003 | 0.024 | 0.087 | 0.184 | 0.234 | 0.154 | 0.028 | 0.003 |
| 10                      | 0.000 | 0.000 | 0.000 | 0.001 | 0.006 | 0.035 | 0.111 | 0.218 | 0.246 | 0.100 | 0.021 |
| 11                      | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.010 | 0.045 | 0.139 | 0.268 | 0.245 | 0.111 |
| 12                      | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.011 | 0.054 | 0.179 | 0.367 | 0.351 |
| 13                      | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.010 | 0.055 | 0.254 | 0.513 |

(*continued*)

## B.1 Binomial Probability Distribution (*concluded*)

| $n = 14$    |       |       |       |       |       |       |       |       |       |       |       |
|-------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Probability |       |       |       |       |       |       |       |       |       |       |       |
| $x$         | 0.05  | 0.10  | 0.20  | 0.30  | 0.40  | 0.50  | 0.60  | 0.70  | 0.80  | 0.90  | 0.95  |
| 0           | 0.488 | 0.229 | 0.044 | 0.007 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 1           | 0.359 | 0.356 | 0.154 | 0.041 | 0.007 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2           | 0.123 | 0.257 | 0.250 | 0.113 | 0.032 | 0.006 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 |
| 3           | 0.026 | 0.114 | 0.250 | 0.194 | 0.085 | 0.022 | 0.003 | 0.000 | 0.000 | 0.000 | 0.000 |
| 4           | 0.004 | 0.035 | 0.172 | 0.229 | 0.155 | 0.061 | 0.014 | 0.001 | 0.000 | 0.000 | 0.000 |
| 5           | 0.000 | 0.008 | 0.086 | 0.196 | 0.207 | 0.122 | 0.041 | 0.007 | 0.000 | 0.000 | 0.000 |
| 6           | 0.000 | 0.001 | 0.032 | 0.126 | 0.207 | 0.183 | 0.092 | 0.023 | 0.002 | 0.000 | 0.000 |
| 7           | 0.000 | 0.000 | 0.009 | 0.062 | 0.157 | 0.209 | 0.157 | 0.062 | 0.009 | 0.000 | 0.000 |
| 8           | 0.000 | 0.000 | 0.002 | 0.023 | 0.092 | 0.183 | 0.207 | 0.126 | 0.032 | 0.001 | 0.000 |
| 9           | 0.000 | 0.000 | 0.000 | 0.007 | 0.041 | 0.122 | 0.207 | 0.196 | 0.086 | 0.008 | 0.000 |
| 10          | 0.000 | 0.000 | 0.000 | 0.001 | 0.014 | 0.061 | 0.155 | 0.229 | 0.172 | 0.035 | 0.004 |
| 11          | 0.000 | 0.000 | 0.000 | 0.000 | 0.003 | 0.022 | 0.085 | 0.194 | 0.250 | 0.114 | 0.026 |
| 12          | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.006 | 0.032 | 0.113 | 0.250 | 0.257 | 0.123 |
| 13          | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.007 | 0.041 | 0.154 | 0.356 | 0.359 |
| 14          | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.007 | 0.044 | 0.229 | 0.488 |

| $n = 15$    |       |       |       |       |       |       |       |       |       |       |       |
|-------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Probability |       |       |       |       |       |       |       |       |       |       |       |
| $x$         | 0.05  | 0.10  | 0.20  | 0.30  | 0.40  | 0.50  | 0.60  | 0.70  | 0.80  | 0.90  | 0.95  |
| 0           | 0.463 | 0.206 | 0.035 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 1           | 0.366 | 0.343 | 0.132 | 0.031 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2           | 0.135 | 0.267 | 0.231 | 0.092 | 0.022 | 0.003 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 3           | 0.031 | 0.129 | 0.250 | 0.170 | 0.063 | 0.014 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 |
| 4           | 0.005 | 0.043 | 0.188 | 0.219 | 0.127 | 0.042 | 0.007 | 0.001 | 0.000 | 0.000 | 0.000 |
| 5           | 0.001 | 0.010 | 0.103 | 0.206 | 0.186 | 0.092 | 0.024 | 0.003 | 0.000 | 0.000 | 0.000 |
| 6           | 0.000 | 0.002 | 0.043 | 0.147 | 0.207 | 0.153 | 0.061 | 0.012 | 0.001 | 0.000 | 0.000 |
| 7           | 0.000 | 0.000 | 0.014 | 0.081 | 0.177 | 0.196 | 0.118 | 0.035 | 0.003 | 0.000 | 0.000 |
| 8           | 0.000 | 0.000 | 0.003 | 0.035 | 0.118 | 0.196 | 0.177 | 0.081 | 0.014 | 0.000 | 0.000 |
| 9           | 0.000 | 0.000 | 0.001 | 0.012 | 0.061 | 0.153 | 0.207 | 0.147 | 0.043 | 0.002 | 0.000 |
| 10          | 0.000 | 0.000 | 0.000 | 0.003 | 0.024 | 0.092 | 0.186 | 0.206 | 0.103 | 0.010 | 0.001 |
| 11          | 0.000 | 0.000 | 0.000 | 0.001 | 0.007 | 0.042 | 0.127 | 0.219 | 0.188 | 0.043 | 0.005 |
| 12          | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.014 | 0.063 | 0.170 | 0.250 | 0.129 | 0.031 |
| 13          | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.003 | 0.022 | 0.092 | 0.231 | 0.267 | 0.135 |
| 14          | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.005 | 0.031 | 0.132 | 0.343 | 0.366 |
| 15          | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.005 | 0.035 | 0.206 | 0.463 |

## B.2 Poisson Distribution

$\mu$

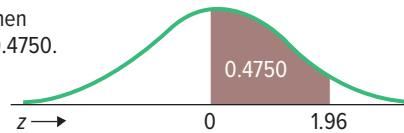
| $x$ | 0.1    | 0.2    | 0.3    | 0.4    | 0.5    | 0.6    | 0.7    | 0.8    | 0.9    |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0   | 0.9048 | 0.8187 | 0.7408 | 0.6703 | 0.6065 | 0.5488 | 0.4966 | 0.4493 | 0.4066 |
| 1   | 0.0905 | 0.1637 | 0.2222 | 0.2681 | 0.3033 | 0.3293 | 0.3476 | 0.3595 | 0.3659 |
| 2   | 0.0045 | 0.0164 | 0.0333 | 0.0536 | 0.0758 | 0.0988 | 0.1217 | 0.1438 | 0.1647 |
| 3   | 0.0002 | 0.0011 | 0.0033 | 0.0072 | 0.0126 | 0.0198 | 0.0284 | 0.0383 | 0.0494 |
| 4   | 0.0000 | 0.0001 | 0.0003 | 0.0007 | 0.0016 | 0.0030 | 0.0050 | 0.0077 | 0.0111 |
| 5   | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0002 | 0.0004 | 0.0007 | 0.0012 | 0.0020 |
| 6   | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0002 | 0.0003 |
| 7   | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

$\mu$

| $x$ | 1.0    | 2.0    | 3.0    | 4.0    | 5.0    | 6.0    | 7.0    | 8.0    | 9.0    |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0   | 0.3679 | 0.1353 | 0.0498 | 0.0183 | 0.0067 | 0.0025 | 0.0009 | 0.0003 | 0.0001 |
| 1   | 0.3679 | 0.2707 | 0.1494 | 0.0733 | 0.0337 | 0.0149 | 0.0064 | 0.0027 | 0.0011 |
| 2   | 0.1839 | 0.2707 | 0.2240 | 0.1465 | 0.0842 | 0.0446 | 0.0223 | 0.0107 | 0.0050 |
| 3   | 0.0613 | 0.1804 | 0.2240 | 0.1954 | 0.1404 | 0.0892 | 0.0521 | 0.0286 | 0.0150 |
| 4   | 0.0153 | 0.0902 | 0.1680 | 0.1954 | 0.1755 | 0.1339 | 0.0912 | 0.0573 | 0.0337 |
| 5   | 0.0031 | 0.0361 | 0.1008 | 0.1563 | 0.1755 | 0.1606 | 0.1277 | 0.0916 | 0.0607 |
| 6   | 0.0005 | 0.0120 | 0.0504 | 0.1042 | 0.1462 | 0.1606 | 0.1490 | 0.1221 | 0.0911 |
| 7   | 0.0001 | 0.0034 | 0.0216 | 0.0595 | 0.1044 | 0.1377 | 0.1490 | 0.1396 | 0.1171 |
| 8   | 0.0000 | 0.0009 | 0.0081 | 0.0298 | 0.0653 | 0.1033 | 0.1304 | 0.1396 | 0.1318 |
| 9   | 0.0000 | 0.0002 | 0.0027 | 0.0132 | 0.0363 | 0.0688 | 0.1014 | 0.1241 | 0.1318 |
| 10  | 0.0000 | 0.0000 | 0.0008 | 0.0053 | 0.0181 | 0.0413 | 0.0710 | 0.0993 | 0.1186 |
| 11  | 0.0000 | 0.0000 | 0.0002 | 0.0019 | 0.0082 | 0.0225 | 0.0452 | 0.0722 | 0.0970 |
| 12  | 0.0000 | 0.0000 | 0.0001 | 0.0006 | 0.0034 | 0.0113 | 0.0263 | 0.0481 | 0.0728 |
| 13  | 0.0000 | 0.0000 | 0.0000 | 0.0002 | 0.0013 | 0.0052 | 0.0142 | 0.0296 | 0.0504 |
| 14  | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0005 | 0.0022 | 0.0071 | 0.0169 | 0.0324 |
| 15  | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0002 | 0.0009 | 0.0033 | 0.0090 | 0.0194 |
| 16  | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0003 | 0.0014 | 0.0045 | 0.0109 |
| 17  | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0006 | 0.0021 | 0.0058 |
| 18  | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0002 | 0.0009 | 0.0029 |
| 19  | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0004 | 0.0014 |
| 20  | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0002 | 0.0006 |
| 21  | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0003 |
| 22  | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 |

## B.3 Areas under the Normal Curve

Example:  
If  $z = 1.96$ , then  
 $P(0 \text{ to } z) = 0.4750$ .



| $z$ | 0.00   | 0.01   | 0.02   | 0.03   | 0.04   | 0.05   | 0.06   | 0.07   | 0.08   | 0.09   |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.0 | 0.0000 | 0.0040 | 0.0080 | 0.0120 | 0.0160 | 0.0199 | 0.0239 | 0.0279 | 0.0319 | 0.0359 |
| 0.1 | 0.0398 | 0.0438 | 0.0478 | 0.0517 | 0.0557 | 0.0596 | 0.0636 | 0.0675 | 0.0714 | 0.0753 |
| 0.2 | 0.0793 | 0.0832 | 0.0871 | 0.0910 | 0.0948 | 0.0987 | 0.1026 | 0.1064 | 0.1103 | 0.1141 |
| 0.3 | 0.1179 | 0.1217 | 0.1255 | 0.1293 | 0.1331 | 0.1368 | 0.1406 | 0.1443 | 0.1480 | 0.1517 |
| 0.4 | 0.1554 | 0.1591 | 0.1628 | 0.1664 | 0.1700 | 0.1736 | 0.1772 | 0.1808 | 0.1844 | 0.1879 |
| 0.5 | 0.1915 | 0.1950 | 0.1985 | 0.2019 | 0.2054 | 0.2088 | 0.2123 | 0.2157 | 0.2190 | 0.2224 |
| 0.6 | 0.2257 | 0.2291 | 0.2324 | 0.2357 | 0.2389 | 0.2422 | 0.2454 | 0.2486 | 0.2517 | 0.2549 |
| 0.7 | 0.2580 | 0.2611 | 0.2642 | 0.2673 | 0.2704 | 0.2734 | 0.2764 | 0.2794 | 0.2823 | 0.2852 |
| 0.8 | 0.2881 | 0.2910 | 0.2939 | 0.2967 | 0.2995 | 0.3023 | 0.3051 | 0.3078 | 0.3106 | 0.3133 |
| 0.9 | 0.3159 | 0.3186 | 0.3212 | 0.3238 | 0.3264 | 0.3289 | 0.3315 | 0.3340 | 0.3365 | 0.3389 |
| 1.0 | 0.3413 | 0.3438 | 0.3461 | 0.3485 | 0.3508 | 0.3531 | 0.3554 | 0.3577 | 0.3599 | 0.3621 |
| 1.1 | 0.3643 | 0.3665 | 0.3686 | 0.3708 | 0.3729 | 0.3749 | 0.3770 | 0.3790 | 0.3810 | 0.3830 |
| 1.2 | 0.3849 | 0.3869 | 0.3888 | 0.3907 | 0.3925 | 0.3944 | 0.3962 | 0.3980 | 0.3997 | 0.4015 |
| 1.3 | 0.4032 | 0.4049 | 0.4066 | 0.4082 | 0.4099 | 0.4115 | 0.4131 | 0.4147 | 0.4162 | 0.4177 |
| 1.4 | 0.4192 | 0.4207 | 0.4222 | 0.4236 | 0.4251 | 0.4265 | 0.4279 | 0.4292 | 0.4306 | 0.4319 |
| 1.5 | 0.4332 | 0.4345 | 0.4357 | 0.4370 | 0.4382 | 0.4394 | 0.4406 | 0.4418 | 0.4429 | 0.4441 |
| 1.6 | 0.4452 | 0.4463 | 0.4474 | 0.4484 | 0.4495 | 0.4505 | 0.4515 | 0.4525 | 0.4535 | 0.4545 |
| 1.7 | 0.4554 | 0.4564 | 0.4573 | 0.4582 | 0.4591 | 0.4599 | 0.4608 | 0.4616 | 0.4625 | 0.4633 |
| 1.8 | 0.4641 | 0.4649 | 0.4656 | 0.4664 | 0.4671 | 0.4678 | 0.4686 | 0.4693 | 0.4699 | 0.4706 |
| 1.9 | 0.4713 | 0.4719 | 0.4726 | 0.4732 | 0.4738 | 0.4744 | 0.4750 | 0.4756 | 0.4761 | 0.4767 |
| 2.0 | 0.4772 | 0.4778 | 0.4783 | 0.4788 | 0.4793 | 0.4798 | 0.4803 | 0.4808 | 0.4812 | 0.4817 |
| 2.1 | 0.4821 | 0.4826 | 0.4830 | 0.4834 | 0.4838 | 0.4842 | 0.4846 | 0.4850 | 0.4854 | 0.4857 |
| 2.2 | 0.4861 | 0.4864 | 0.4868 | 0.4871 | 0.4875 | 0.4878 | 0.4881 | 0.4884 | 0.4887 | 0.4890 |
| 2.3 | 0.4893 | 0.4896 | 0.4898 | 0.4901 | 0.4904 | 0.4906 | 0.4909 | 0.4911 | 0.4913 | 0.4916 |
| 2.4 | 0.4918 | 0.4920 | 0.4922 | 0.4925 | 0.4927 | 0.4929 | 0.4931 | 0.4932 | 0.4934 | 0.4936 |
| 2.5 | 0.4938 | 0.4940 | 0.4941 | 0.4943 | 0.4945 | 0.4946 | 0.4948 | 0.4949 | 0.4951 | 0.4952 |
| 2.6 | 0.4953 | 0.4955 | 0.4956 | 0.4957 | 0.4959 | 0.4960 | 0.4961 | 0.4962 | 0.4963 | 0.4964 |
| 2.7 | 0.4965 | 0.4966 | 0.4967 | 0.4968 | 0.4969 | 0.4970 | 0.4971 | 0.4972 | 0.4973 | 0.4974 |
| 2.8 | 0.4974 | 0.4975 | 0.4976 | 0.4977 | 0.4977 | 0.4978 | 0.4979 | 0.4979 | 0.4980 | 0.4981 |
| 2.9 | 0.4981 | 0.4982 | 0.4982 | 0.4983 | 0.4984 | 0.4984 | 0.4985 | 0.4985 | 0.4986 | 0.4986 |
| 3.0 | 0.4987 | 0.4987 | 0.4987 | 0.4988 | 0.4988 | 0.4989 | 0.4989 | 0.4989 | 0.4990 | 0.4990 |



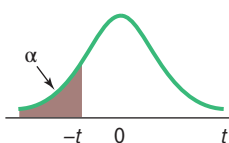
## B.4 Table of Random Numbers

|       |       |       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 02711 | 08182 | 75997 | 79866 | 58095 | 83319 | 80295 | 79741 | 74599 | 84379 |
| 94873 | 90935 | 31684 | 63952 | 09865 | 14491 | 99518 | 93394 | 34691 | 14985 |
| 54921 | 78680 | 06635 | 98689 | 17306 | 25170 | 65928 | 87709 | 30533 | 89736 |
| 77640 | 97636 | 37397 | 93379 | 56454 | 59818 | 45827 | 74164 | 71666 | 46977 |
| 61545 | 00835 | 93251 | 87203 | 36759 | 49197 | 85967 | 01704 | 19634 | 21898 |
| 17147 | 19519 | 22497 | 16857 | 42426 | 84822 | 92598 | 49186 | 88247 | 39967 |
| 13748 | 04742 | 92460 | 85801 | 53444 | 65626 | 58710 | 55406 | 17173 | 69776 |
| 87455 | 14813 | 50373 | 28037 | 91182 | 32786 | 65261 | 11173 | 34376 | 36408 |
| 08999 | 57409 | 91185 | 10200 | 61411 | 23392 | 47797 | 56377 | 71635 | 08601 |
| 78804 | 81333 | 53809 | 32471 | 46034 | 36306 | 22498 | 19239 | 85428 | 55721 |
| 82173 | 26921 | 28472 | 98958 | 07960 | 66124 | 89731 | 95069 | 18625 | 92405 |
| 97594 | 25168 | 89178 | 68190 | 05043 | 17407 | 48201 | 83917 | 11413 | 72920 |
| 73881 | 67176 | 93504 | 42636 | 38233 | 16154 | 96451 | 57925 | 29667 | 30859 |
| 46071 | 22912 | 90326 | 42453 | 88108 | 72064 | 58601 | 32357 | 90610 | 32921 |
| 44492 | 19686 | 12495 | 93135 | 95185 | 77799 | 52441 | 88272 | 22024 | 80631 |
| 31864 | 72170 | 37722 | 55794 | 14636 | 05148 | 54505 | 50113 | 21119 | 25228 |
| 51574 | 90692 | 43339 | 65689 | 76539 | 27909 | 05467 | 21727 | 51141 | 72949 |
| 35350 | 76132 | 92925 | 92124 | 92634 | 35681 | 43690 | 89136 | 35599 | 84138 |
| 46943 | 36502 | 01172 | 46045 | 46991 | 33804 | 80006 | 35542 | 61056 | 75666 |
| 22665 | 87226 | 33304 | 57975 | 03985 | 21566 | 65796 | 72915 | 81466 | 89205 |
| 39437 | 97957 | 11838 | 10433 | 21564 | 51570 | 73558 | 27495 | 34533 | 57808 |
| 77082 | 47784 | 40098 | 97962 | 89845 | 28392 | 78187 | 06112 | 08169 | 11261 |
| 24544 | 25649 | 43370 | 28007 | 06779 | 72402 | 62632 | 53956 | 24709 | 06978 |
| 27503 | 15558 | 37738 | 24849 | 70722 | 71859 | 83736 | 06016 | 94397 | 12529 |
| 24590 | 24545 | 06435 | 52758 | 45685 | 90151 | 46516 | 49644 | 92686 | 84870 |
| 48155 | 86226 | 40359 | 28723 | 15364 | 69125 | 12609 | 57171 | 86857 | 31702 |
| 20226 | 53752 | 90648 | 24362 | 83314 | 00014 | 19207 | 69413 | 97016 | 86290 |
| 70178 | 73444 | 38790 | 53626 | 93780 | 18629 | 68766 | 24371 | 74639 | 30782 |
| 10169 | 41465 | 51935 | 05711 | 09799 | 79077 | 88159 | 33437 | 68519 | 03040 |
| 81084 | 03701 | 28598 | 70013 | 63794 | 53169 | 97054 | 60303 | 23259 | 96196 |
| 69202 | 20777 | 21727 | 81511 | 51887 | 16175 | 53746 | 46516 | 70339 | 62727 |
| 80561 | 95787 | 89426 | 93325 | 86412 | 57479 | 54194 | 52153 | 19197 | 81877 |
| 08199 | 26703 | 95128 | 48599 | 09333 | 12584 | 24374 | 31232 | 61782 | 44032 |
| 98883 | 28220 | 39358 | 53720 | 80161 | 83371 | 15181 | 11131 | 12219 | 55920 |
| 84568 | 69286 | 76054 | 21615 | 80883 | 36797 | 82845 | 39139 | 90900 | 18172 |
| 04269 | 35173 | 95745 | 53893 | 86022 | 77722 | 52498 | 84193 | 22448 | 22571 |
| 10538 | 13124 | 36099 | 13140 | 37706 | 44562 | 57179 | 44693 | 67877 | 01549 |
| 77843 | 24955 | 25900 | 63843 | 95029 | 93859 | 93634 | 20205 | 66294 | 41218 |
| 12034 | 94636 | 49455 | 76362 | 83532 | 31062 | 69903 | 91186 | 65768 | 55949 |
| 10524 | 72829 | 47641 | 93315 | 80875 | 28090 | 97728 | 52560 | 34937 | 79548 |
| 68935 | 76632 | 46984 | 61772 | 92786 | 22651 | 07086 | 89754 | 44143 | 97687 |
| 89450 | 65665 | 29190 | 43709 | 11172 | 34481 | 95977 | 47535 | 25658 | 73898 |
| 90696 | 20451 | 24211 | 97310 | 60446 | 73530 | 62865 | 96574 | 13829 | 72226 |
| 49006 | 32047 | 93086 | 00112 | 20470 | 17136 | 28255 | 86328 | 07293 | 38809 |
| 74591 | 87025 | 52368 | 59416 | 34417 | 70557 | 86746 | 55809 | 53628 | 12000 |
| 06315 | 17012 | 77103 | 00968 | 07235 | 10728 | 42189 | 33292 | 51487 | 64443 |
| 62386 | 09184 | 62092 | 46617 | 99419 | 64230 | 95034 | 85481 | 07857 | 42510 |
| 86848 | 82122 | 04028 | 36959 | 87827 | 12813 | 08627 | 80699 | 13345 | 51695 |
| 65643 | 69480 | 46598 | 04501 | 40403 | 91408 | 32343 | 48130 | 49303 | 90689 |
| 11084 | 46534 | 78957 | 77353 | 39578 | 77868 | 22970 | 84349 | 09184 | 70603 |

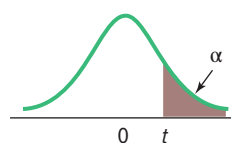
## B.5 Student's *t* Distribution



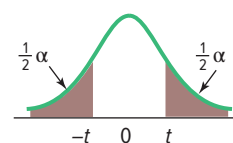
Confidence interval



Left-tailed test



Right-tailed test



Two-tailed test

| Confidence Intervals, <i>c</i> |   |       |        |        |        |         |
|--------------------------------|---|-------|--------|--------|--------|---------|
| <i>df</i>                      | 80%   | 90%   | 95%    | 98%    | 99%    | 99.9%   |
|                                | Level of Significance for One-Tailed Test, $\alpha$ |       |        |        |        |         |
|                                | 0.10  | 0.05  | 0.025  | 0.01   | 0.005  | 0.0005  |
|                                | Level of Significance for Two-Tailed Test, $\alpha$ |       |        |        |        |         |
|                                | 0.20  | 0.10  | 0.05   | 0.02   | 0.01   | 0.001   |
| 1                              | 3.078   | 6.314 | 12.706 | 31.821 | 63.657 | 636.619 |
| 2                              | 1.886   | 2.920 | 4.303  | 6.965  | 9.925  | 31.599  |
| 3                              | 1.638   | 2.353 | 3.182  | 4.541  | 5.841  | 12.924  |
| 4                              | 1.533   | 2.132 | 2.776  | 3.747  | 4.604  | 8.610   |
| 5                              | 1.476   | 2.015 | 2.571  | 3.365  | 4.032  | 6.869   |
| 6                              | 1.440   | 1.943 | 2.447  | 3.143  | 3.707  | 5.959   |
| 7                              | 1.415   | 1.895 | 2.365  | 2.998  | 3.499  | 5.408   |
| 8                              | 1.397   | 1.860 | 2.306  | 2.896  | 3.355  | 5.041   |
| 9                              | 1.383   | 1.833 | 2.262  | 2.821  | 3.250  | 4.781   |
| 10                             | 1.372   | 1.812 | 2.228  | 2.764  | 3.169  | 4.587   |
| 11                             | 1.363   | 1.796 | 2.201  | 2.718  | 3.106  | 4.437   |
| 12                             | 1.356   | 1.782 | 2.179  | 2.681  | 3.055  | 4.318   |
| 13                             | 1.350   | 1.771 | 2.160  | 2.650  | 3.012  | 4.221   |
| 14                             | 1.345   | 1.761 | 2.145  | 2.624  | 2.977  | 4.140   |
| 15                             | 1.341   | 1.753 | 2.131  | 2.602  | 2.947  | 4.073   |
| 16                             | 1.337   | 1.746 | 2.120  | 2.583  | 2.921  | 4.015   |
| 17                             | 1.333   | 1.740 | 2.110  | 2.567  | 2.898  | 3.965   |
| 18                             | 1.330   | 1.734 | 2.101  | 2.552  | 2.878  | 3.922   |
| 19                             | 1.328   | 1.729 | 2.093  | 2.539  | 2.861  | 3.883   |
| 20                             | 1.325   | 1.725 | 2.086  | 2.528  | 2.845  | 3.850   |
| 21                             | 1.323   | 1.721 | 2.080  | 2.518  | 2.831  | 3.819   |
| 22                             | 1.321   | 1.717 | 2.074  | 2.508  | 2.819  | 3.792   |
| 23                             | 1.319   | 1.714 | 2.069  | 2.500  | 2.807  | 3.768   |
| 24                             | 1.318   | 1.711 | 2.064  | 2.492  | 2.797  | 3.745   |
| 25                             | 1.316   | 1.708 | 2.060  | 2.485  | 2.787  | 3.725   |
| 26                             | 1.315   | 1.706 | 2.056  | 2.479  | 2.779  | 3.707   |
| 27                             | 1.314   | 1.703 | 2.052  | 2.473  | 2.771  | 3.690   |
| 28                             | 1.313   | 1.701 | 2.048  | 2.467  | 2.763  | 3.674   |
| 29                             | 1.311   | 1.699 | 2.045  | 2.462  | 2.756  | 3.659   |
| 30                             | 1.310   | 1.697 | 2.042  | 2.457  | 2.750  | 3.646   |
| 31                             | 1.309   | 1.696 | 2.040  | 2.453  | 2.744  | 3.633   |
| 32                             | 1.309   | 1.694 | 2.037  | 2.449  | 2.738  | 3.622   |
| 33                             | 1.308   | 1.692 | 2.035  | 2.445  | 2.733  | 3.611   |
| 34                             | 1.307   | 1.691 | 2.032  | 2.441  | 2.728  | 3.601   |
| 35                             | 1.306   | 1.690 | 2.030  | 2.438  | 2.724  | 3.591   |

| Confidence Intervals, <i>c</i> |   |       |       |       |       |        |
|--------------------------------|---|-------|-------|-------|-------|--------|
| <i>df</i>                      | 80%   | 90%   | 95%   | 98%   | 99%   | 99.9%  |
|                                | Level of Significance for One-Tailed Test, $\alpha$ |       |       |       |       |        |
|                                | 0.10  | 0.05  | 0.025 | 0.01  | 0.005 | 0.0005 |
|                                | Level of Significance for Two-Tailed Test, $\alpha$ |       |       |       |       |        |
|                                | 0.20  | 0.10  | 0.05  | 0.02  | 0.01  | 0.001  |
| 36                             | 1.306   | 1.688 | 2.028 | 2.434 | 2.719 | 3.582  |
| 37                             | 1.305   | 1.687 | 2.026 | 2.431 | 2.715 | 3.574  |
| 38                             | 1.304   | 1.686 | 2.024 | 2.429 | 2.712 | 3.566  |
| 39                             | 1.304   | 1.685 | 2.023 | 2.426 | 2.708 | 3.558  |
| 40                             | 1.303   | 1.684 | 2.021 | 2.423 | 2.704 | 3.551  |
| 41                             | 1.303   | 1.683 | 2.020 | 2.421 | 2.701 | 3.544  |
| 42                             | 1.302   | 1.682 | 2.018 | 2.418 | 2.698 | 3.538  |
| 43                             | 1.302   | 1.681 | 2.017 | 2.416 | 2.695 | 3.532  |
| 44                             | 1.301   | 1.680 | 2.015 | 2.414 | 2.692 | 3.526  |
| 45                             | 1.301   | 1.679 | 2.014 | 2.412 | 2.690 | 3.520  |
| 46                             | 1.300   | 1.679 | 2.013 | 2.410 | 2.687 | 3.515  |
| 47                             | 1.300   | 1.678 | 2.012 | 2.408 | 2.685 | 3.510  |
| 48                             | 1.299   | 1.677 | 2.011 | 2.407 | 2.682 | 3.505  |
| 49                             | 1.299   | 1.677 | 2.010 | 2.405 | 2.680 | 3.500  |
| 50                             | 1.299   | 1.676 | 2.009 | 2.403 | 2.678 | 3.496  |
| 51                             | 1.298   | 1.675 | 2.008 | 2.402 | 2.676 | 3.492  |
| 52                             | 1.298   | 1.675 | 2.007 | 2.400 | 2.674 | 3.488  |
| 53                             | 1.298   | 1.674 | 2.006 | 2.399 | 2.672 | 3.484  |
| 54                             | 1.297   | 1.674 | 2.005 | 2.397 | 2.670 | 3.480  |
| 55                             | 1.297   | 1.673 | 2.004 | 2.396 | 2.668 | 3.476  |
| 56                             | 1.297   | 1.673 | 2.003 | 2.395 | 2.667 | 3.473  |
| 57                             | 1.297   | 1.672 | 2.002 | 2.394 | 2.665 | 3.470  |
| 58                             | 1.296   | 1.672 | 2.002 | 2.392 | 2.663 | 3.466  |
| 59                             | 1.296   | 1.671 | 2.001 | 2.391 | 2.662 | 3.463  |
| 60                             | 1.296   | 1.671 | 2.000 | 2.390 | 2.660 | 3.460  |
| 61                             | 1.296   | 1.670 | 2.000 | 2.389 | 2.659 | 3.457  |
| 62                             | 1.295   | 1.670 | 1.999 | 2.388 | 2.657 | 3.454  |
| 63                             | 1.295   | 1.669 | 1.998 | 2.387 | 2.656 | 3.452  |
| 64                             | 1.295   | 1.669 | 1.998 | 2.386 | 2.655 | 3.449  |
| 65                             | 1.295   | 1.669 | 1.997 | 2.385 | 2.654 | 3.447  |
| 66                             | 1.295   | 1.668 | 1.997 | 2.384 | 2.652 | 3.444  |
| 67                             | 1.294   | 1.668 | 1.996 | 2.383 | 2.651 | 3.442  |
| 68                             | 1.294   | 1.668 | 1.995 | 2.382 | 2.650 | 3.439  |
| 69                             | 1.294   | 1.667 | 1.995 | 2.382 | 2.649 | 3.437  |
| 70                             | 1.294   | 1.667 | 1.994 | 2.381 | 2.648 | 3.435  |

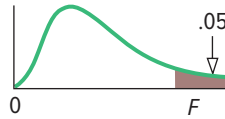
(continued)

## B.5 Student's *t* Distribution (concluded)

| Confidence Intervals, <i>c</i> |   |       |       |       |       |        |
|--------------------------------|---|-------|-------|-------|-------|--------|
|                                | 80%   | 90%   | 95%   | 98%   | 99%   | 99.9%  |
| <i>df</i>                      | Level of Significance for One-Tailed Test, $\alpha$ |       |       |       |       |        |
|                                | 0.10  | 0.05  | 0.025 | 0.01  | 0.005 | 0.0005 |
|                                | Level of Significance for Two-Tailed Test, $\alpha$ |       |       |       |       |        |
|                                | 0.20  | 0.10  | 0.05  | 0.02  | 0.01  | 0.001  |
| 71                             | 1.294   | 1.667 | 1.994 | 2.380 | 2.647 | 3.433  |
| 72                             | 1.293   | 1.666 | 1.993 | 2.379 | 2.646 | 3.431  |
| 73                             | 1.293   | 1.666 | 1.993 | 2.379 | 2.645 | 3.429  |
| 74                             | 1.293   | 1.666 | 1.993 | 2.378 | 2.644 | 3.427  |
| 75                             | 1.293   | 1.665 | 1.992 | 2.377 | 2.643 | 3.425  |
| 76                             | 1.293   | 1.665 | 1.992 | 2.376 | 2.642 | 3.423  |
| 77                             | 1.293   | 1.665 | 1.991 | 2.376 | 2.641 | 3.421  |
| 78                             | 1.292   | 1.665 | 1.991 | 2.375 | 2.640 | 3.420  |
| 79                             | 1.292   | 1.664 | 1.990 | 2.374 | 2.640 | 3.418  |
| 80                             | 1.292   | 1.664 | 1.990 | 2.374 | 2.639 | 3.416  |
| 81                             | 1.292   | 1.664 | 1.990 | 2.373 | 2.638 | 3.415  |
| 82                             | 1.292   | 1.664 | 1.989 | 2.373 | 2.637 | 3.413  |
| 83                             | 1.292   | 1.663 | 1.989 | 2.372 | 2.636 | 3.412  |
| 84                             | 1.292   | 1.663 | 1.989 | 2.372 | 2.636 | 3.410  |
| 85                             | 1.292   | 1.663 | 1.988 | 2.371 | 2.635 | 3.409  |
| 86                             | 1.291   | 1.663 | 1.988 | 2.370 | 2.634 | 3.407  |
| 87                             | 1.291   | 1.663 | 1.988 | 2.370 | 2.634 | 3.406  |
| 88                             | 1.291   | 1.662 | 1.987 | 2.369 | 2.633 | 3.405  |

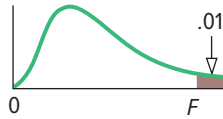
| Confidence Intervals, <i>c</i> |   |       |       |       |       |        |
|--------------------------------|---|-------|-------|-------|-------|--------|
|                                | 80%   | 90%   | 95%   | 98%   | 99%   | 99.9%  |
| <i>df</i>                      | Level of Significance for One-Tailed Test, $\alpha$ |       |       |       |       |        |
|                                | 0.10  | 0.05  | 0.025 | 0.01  | 0.005 | 0.0005 |
|                                | Level of Significance for Two-Tailed Test, $\alpha$ |       |       |       |       |        |
|                                | 0.20  | 0.10  | 0.05  | 0.02  | 0.01  | 0.001  |
| 89                             | 1.291   | 1.662 | 1.987 | 2.369 | 2.632 | 3.403  |
| 90                             | 1.291   | 1.662 | 1.987 | 2.368 | 2.632 | 3.402  |
| 91                             | 1.291   | 1.662 | 1.986 | 2.368 | 2.631 | 3.401  |
| 92                             | 1.291   | 1.662 | 1.986 | 2.368 | 2.630 | 3.399  |
| 93                             | 1.291   | 1.661 | 1.986 | 2.367 | 2.630 | 3.398  |
| 94                             | 1.291   | 1.661 | 1.986 | 2.367 | 2.629 | 3.397  |
| 95                             | 1.291   | 1.661 | 1.985 | 2.366 | 2.629 | 3.396  |
| 96                             | 1.290   | 1.661 | 1.985 | 2.366 | 2.628 | 3.395  |
| 97                             | 1.290   | 1.661 | 1.985 | 2.365 | 2.627 | 3.394  |
| 98                             | 1.290   | 1.661 | 1.984 | 2.365 | 2.627 | 3.393  |
| 99                             | 1.290   | 1.660 | 1.984 | 2.365 | 2.626 | 3.392  |
| 100                            | 1.290   | 1.660 | 1.984 | 2.364 | 2.626 | 3.390  |
| 120                            | 1.289   | 1.658 | 1.980 | 2.358 | 2.617 | 3.373  |
| 140                            | 1.288   | 1.656 | 1.977 | 2.353 | 2.611 | 3.361  |
| 160                            | 1.287   | 1.654 | 1.975 | 2.350 | 2.607 | 3.352  |
| 180                            | 1.286   | 1.653 | 1.973 | 2.347 | 2.603 | 3.345  |
| 200                            | 1.286   | 1.653 | 1.972 | 2.345 | 2.601 | 3.340  |
| $\infty$                       | 1.282   | 1.645 | 1.960 | 2.326 | 2.576 | 3.291  |

## B.6A Critical Values of the $F$ Distribution ( $\alpha = .05$ )



|  |      | Degrees of Freedom for the Numerator |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
|--|------|--------------------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
|  |      | 1                                    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   | 12   | 15   | 20   | 24   | 30   | 40   |      |
| Degrees of Freedom for the Denominator | 1    | 161                                  | 200  | 216  | 225  | 230  | 234  | 237  | 239  | 241  | 242  | 244  | 246  | 248  | 249  | 250  | 251  |      |
|  | 2    | 18.5                                 | 19.0 | 19.2 | 19.2 | 19.3 | 19.3 | 19.4 | 19.4 | 19.4 | 19.4 | 19.4 | 19.4 | 19.4 | 19.4 | 19.5 | 19.5 | 19.5 |
|  | 3    | 10.1                                 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 | 8.79 | 8.74 | 8.70 | 8.66 | 8.64 | 8.62 | 8.59 | 8.59 |
|  | 4    | 7.71                                 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 | 5.96 | 5.91 | 5.86 | 5.80 | 5.77 | 5.75 | 5.72 | 5.72 |
|  | 5    | 6.61                                 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 | 4.74 | 4.68 | 4.62 | 4.56 | 4.53 | 4.50 | 4.46 | 4.46 |
|  | 6    | 5.99                                 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 | 4.06 | 4.00 | 3.94 | 3.87 | 3.84 | 3.81 | 3.77 | 3.77 |
|  | 7    | 5.59                                 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 | 3.64 | 3.57 | 3.51 | 3.44 | 3.41 | 3.38 | 3.34 | 3.34 |
|  | 8    | 5.32                                 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 | 3.35 | 3.28 | 3.22 | 3.15 | 3.12 | 3.08 | 3.04 | 3.04 |
|  | 9    | 5.12                                 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 | 3.14 | 3.07 | 3.01 | 2.94 | 2.90 | 2.86 | 2.83 | 2.83 |
|  | 10   | 4.96                                 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.98 | 2.91 | 2.85 | 2.77 | 2.74 | 2.70 | 2.66 | 2.66 |
|  | 11   | 4.84                                 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 | 2.85 | 2.79 | 2.72 | 2.65 | 2.61 | 2.57 | 2.53 | 2.53 |
|  | 12   | 4.75                                 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 | 2.75 | 2.69 | 2.62 | 2.54 | 2.51 | 2.47 | 2.43 | 2.43 |
|  | 13   | 4.67                                 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 | 2.67 | 2.60 | 2.53 | 2.46 | 2.42 | 2.38 | 2.34 | 2.34 |
|  | 14   | 4.60                                 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 | 2.60 | 2.53 | 2.46 | 2.39 | 2.35 | 2.31 | 2.27 | 2.27 |
|  | 15   | 4.54                                 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 | 2.54 | 2.48 | 2.40 | 2.33 | 2.29 | 2.25 | 2.20 | 2.20 |
|  | 16   | 4.49                                 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 | 2.49 | 2.42 | 2.35 | 2.28 | 2.24 | 2.19 | 2.15 | 2.15 |
|  | 17   | 4.45                                 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.49 | 2.45 | 2.38 | 2.31 | 2.23 | 2.19 | 2.15 | 2.10 | 2.10 |
|  | 18   | 4.41                                 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 | 2.41 | 2.34 | 2.27 | 2.19 | 2.15 | 2.11 | 2.06 | 2.06 |
|  | 19   | 4.38                                 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2.48 | 2.42 | 2.38 | 2.31 | 2.23 | 2.16 | 2.11 | 2.07 | 2.03 | 2.03 |
|  | 20   | 4.35                                 | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.51 | 2.45 | 2.39 | 2.35 | 2.28 | 2.20 | 2.12 | 2.08 | 2.04 | 1.99 | 1.99 |
|  | 21   | 4.32                                 | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.49 | 2.42 | 2.37 | 2.32 | 2.25 | 2.18 | 2.10 | 2.05 | 2.01 | 1.96 | 1.96 |
|  | 22   | 4.30                                 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.46 | 2.40 | 2.34 | 2.30 | 2.23 | 2.15 | 2.07 | 2.03 | 1.98 | 1.94 | 1.94 |
|  | 23   | 4.28                                 | 3.42 | 3.03 | 2.80 | 2.64 | 2.53 | 2.44 | 2.37 | 2.32 | 2.27 | 2.20 | 2.13 | 2.05 | 2.01 | 1.96 | 1.91 | 1.91 |
|  | 24   | 4.26                                 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.42 | 2.36 | 2.30 | 2.25 | 2.18 | 2.11 | 2.03 | 1.98 | 1.94 | 1.89 | 1.89 |
|  | 25   | 4.24                                 | 3.39 | 2.99 | 2.76 | 2.60 | 2.49 | 2.40 | 2.34 | 2.28 | 2.24 | 2.16 | 2.09 | 2.01 | 1.96 | 1.92 | 1.87 | 1.87 |
| 30                                     | 4.17 | 3.32                                 | 2.92 | 2.69 | 2.53 | 2.42 | 2.33 | 2.27 | 2.21 | 2.16 | 2.09 | 2.01 | 1.93 | 1.89 | 1.84 | 1.79 | 1.79 |      |
| 40                                     | 4.08 | 3.23                                 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2.12 | 2.08 | 2.00 | 1.92 | 1.84 | 1.79 | 1.74 | 1.69 | 1.69 |      |
| 60                                     | 4.00 | 3.15                                 | 2.76 | 2.53 | 2.37 | 2.25 | 2.17 | 2.10 | 2.04 | 1.99 | 1.92 | 1.84 | 1.75 | 1.70 | 1.65 | 1.59 | 1.59 |      |
| 120                                    | 3.92 | 3.07                                 | 2.68 | 2.45 | 2.29 | 2.18 | 2.09 | 2.02 | 1.96 | 1.91 | 1.83 | 1.75 | 1.66 | 1.61 | 1.55 | 1.50 | 1.50 |      |
| $\infty$                               | 3.84 | 3.00                                 | 2.60 | 2.37 | 2.21 | 2.10 | 2.01 | 1.94 | 1.88 | 1.83 | 1.75 | 1.67 | 1.57 | 1.52 | 1.46 | 1.39 | 1.39 |      |

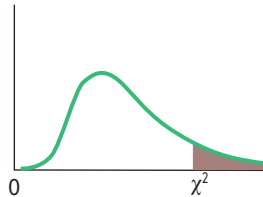
## B.6B Critical Values of the F Distribution ( $\alpha = .01$ )



|  |      | Degrees of Freedom for the Numerator |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
|--|------|--------------------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
|  |      | 1                                    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   | 12   | 15   | 20   | 24   | 30   | 40   |
| Degrees of Freedom for the Denominator | 1    | 4052                                 | 5000 | 5403 | 5625 | 5764 | 5859 | 5928 | 5981 | 6022 | 6056 | 6106 | 6157 | 6209 | 6235 | 6261 | 6287 |
|  | 2    | 98.5                                 | 99.0 | 99.2 | 99.2 | 99.3 | 99.3 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 99.5 | 99.5 | 99.5 |
|  | 3    | 34.1                                 | 30.8 | 29.5 | 28.7 | 28.2 | 27.9 | 27.7 | 27.5 | 27.3 | 27.2 | 27.1 | 26.9 | 26.7 | 26.6 | 26.5 | 26.4 |
|  | 4    | 21.2                                 | 18.0 | 16.7 | 16.0 | 15.5 | 15.2 | 15.0 | 14.8 | 14.7 | 14.5 | 14.4 | 14.2 | 14.0 | 13.9 | 13.8 | 13.7 |
|  | 5    | 16.3                                 | 13.3 | 12.1 | 11.4 | 11.0 | 10.7 | 10.5 | 10.3 | 10.2 | 10.1 | 9.89 | 9.72 | 9.55 | 9.47 | 9.38 | 9.29 |
|  | 6    | 13.7                                 | 10.9 | 9.78 | 9.15 | 8.75 | 8.47 | 8.26 | 8.10 | 7.98 | 7.87 | 7.72 | 7.56 | 7.40 | 7.31 | 7.23 | 7.14 |
|  | 7    | 12.2                                 | 9.55 | 8.45 | 7.85 | 7.46 | 7.19 | 6.99 | 6.84 | 6.72 | 6.62 | 6.47 | 6.31 | 6.16 | 6.07 | 5.99 | 5.91 |
|  | 8    | 11.3                                 | 8.65 | 7.59 | 7.01 | 6.63 | 6.37 | 6.18 | 6.03 | 5.91 | 5.81 | 5.67 | 5.52 | 5.36 | 5.28 | 5.20 | 5.12 |
|  | 9    | 10.6                                 | 8.02 | 6.99 | 6.42 | 6.06 | 5.80 | 5.61 | 5.47 | 5.35 | 5.26 | 5.11 | 4.96 | 4.81 | 4.73 | 4.65 | 4.57 |
|  | 10   | 10.0                                 | 7.56 | 6.55 | 5.99 | 5.64 | 5.39 | 5.20 | 5.06 | 4.94 | 4.85 | 4.71 | 4.56 | 4.41 | 4.33 | 4.25 | 4.17 |
|  | 11   | 9.65                                 | 7.21 | 6.22 | 5.67 | 5.32 | 5.07 | 4.89 | 4.74 | 4.63 | 4.54 | 4.40 | 4.25 | 4.10 | 4.02 | 3.94 | 3.86 |
|  | 12   | 9.33                                 | 6.93 | 5.95 | 5.41 | 5.06 | 4.82 | 4.64 | 4.50 | 4.39 | 4.30 | 4.16 | 4.01 | 3.86 | 3.78 | 3.70 | 3.62 |
|  | 13   | 9.07                                 | 6.70 | 5.74 | 5.21 | 4.86 | 4.62 | 4.44 | 4.30 | 4.19 | 4.10 | 3.96 | 3.82 | 3.66 | 3.59 | 3.51 | 3.43 |
|  | 14   | 8.86                                 | 6.51 | 5.56 | 5.04 | 4.69 | 4.46 | 4.28 | 4.14 | 4.03 | 3.94 | 3.80 | 3.66 | 3.51 | 3.43 | 3.35 | 3.27 |
|  | 15   | 8.68                                 | 6.36 | 5.42 | 4.89 | 4.56 | 4.32 | 4.14 | 4.00 | 3.89 | 3.80 | 3.67 | 3.52 | 3.37 | 3.29 | 3.21 | 3.13 |
|  | 16   | 8.53                                 | 6.23 | 5.29 | 4.77 | 4.44 | 4.20 | 4.03 | 3.89 | 3.78 | 3.69 | 3.55 | 3.41 | 3.26 | 3.18 | 3.10 | 3.02 |
|  | 17   | 8.40                                 | 6.11 | 5.18 | 4.67 | 4.34 | 4.10 | 3.93 | 3.79 | 3.68 | 3.59 | 3.46 | 3.31 | 3.16 | 3.08 | 3.00 | 2.92 |
|  | 18   | 8.29                                 | 6.01 | 5.09 | 4.58 | 4.25 | 4.01 | 3.84 | 3.71 | 3.60 | 3.51 | 3.37 | 3.23 | 3.08 | 3.00 | 2.92 | 2.84 |
|  | 19   | 8.18                                 | 5.93 | 5.01 | 4.50 | 4.17 | 3.94 | 3.77 | 3.63 | 3.52 | 3.43 | 3.30 | 3.15 | 3.00 | 2.92 | 2.84 | 2.76 |
|  | 20   | 8.10                                 | 5.85 | 4.94 | 4.43 | 4.10 | 3.87 | 3.70 | 3.56 | 3.46 | 3.37 | 3.23 | 3.09 | 2.94 | 2.86 | 2.78 | 2.69 |
|  | 21   | 8.02                                 | 5.78 | 4.87 | 4.37 | 4.04 | 3.81 | 3.64 | 3.51 | 3.40 | 3.31 | 3.17 | 3.03 | 2.88 | 2.80 | 2.72 | 2.64 |
|  | 22   | 7.95                                 | 5.72 | 4.82 | 4.31 | 3.99 | 3.76 | 3.59 | 3.45 | 3.35 | 3.26 | 3.12 | 2.98 | 2.83 | 2.75 | 2.67 | 2.58 |
|  | 23   | 7.88                                 | 5.66 | 4.76 | 4.26 | 3.94 | 3.71 | 3.54 | 3.41 | 3.30 | 3.21 | 3.07 | 2.93 | 2.78 | 2.70 | 2.62 | 2.54 |
|  | 24   | 7.82                                 | 5.61 | 4.72 | 4.22 | 3.90 | 3.67 | 3.50 | 3.36 | 3.26 | 3.17 | 3.03 | 2.89 | 2.74 | 2.66 | 2.58 | 2.49 |
|  | 25   | 7.77                                 | 5.57 | 4.68 | 4.18 | 3.85 | 3.63 | 3.46 | 3.32 | 3.22 | 3.13 | 2.99 | 2.85 | 2.70 | 2.62 | 2.54 | 2.45 |
| 30                                     | 7.56 | 5.39                                 | 4.51 | 4.02 | 3.70 | 3.47 | 3.30 | 3.17 | 3.07 | 2.98 | 2.84 | 2.70 | 2.55 | 2.47 | 2.39 | 2.30 |      |
| 40                                     | 7.31 | 5.18                                 | 4.31 | 3.83 | 3.51 | 3.29 | 3.12 | 2.99 | 2.89 | 2.80 | 2.66 | 2.52 | 2.37 | 2.29 | 2.20 | 2.11 |      |
| 60                                     | 7.08 | 4.98                                 | 4.13 | 3.65 | 3.34 | 3.12 | 2.95 | 2.82 | 2.72 | 2.63 | 2.50 | 2.35 | 2.20 | 2.12 | 2.03 | 1.94 |      |
| 120                                    | 6.85 | 4.79                                 | 3.95 | 3.48 | 3.17 | 2.96 | 2.79 | 2.66 | 2.56 | 2.47 | 2.34 | 2.19 | 2.03 | 1.95 | 1.86 | 1.76 |      |
| $\infty$                               | 6.63 | 4.61                                 | 3.78 | 3.32 | 3.02 | 2.80 | 2.64 | 2.51 | 2.41 | 2.32 | 2.18 | 2.04 | 1.88 | 1.79 | 1.70 | 1.59 |      |

## B.7 Critical Values of Chi-Square

This table contains the values of  $\chi^2$  that correspond to a specific right-tail area and specific number of degrees of freedom.



Example: With 17 *df* and a .02 area in the upper tail,  $\chi^2 = 30.995$ .

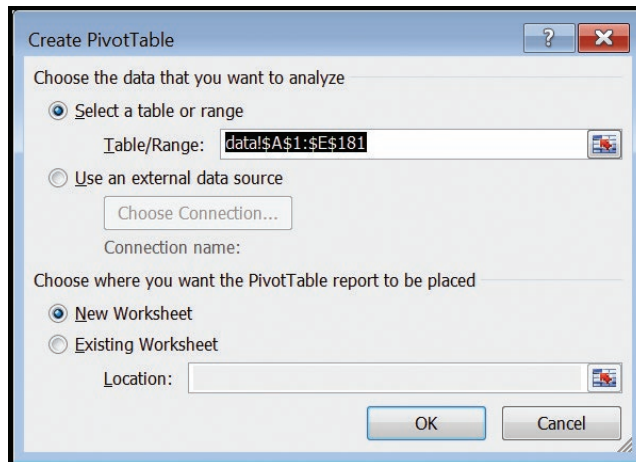
| Degrees of Freedom, <i>df</i> | Right-Tail Area |        |        |        |
|-------------------------------|-----------------|--------|--------|--------|
|                               | 0.10            | 0.05   | 0.02   | 0.01   |
| 1                             | 2.706           | 3.841  | 5.412  | 6.635  |
| 2                             | 4.605           | 5.991  | 7.824  | 9.210  |
| 3                             | 6.251           | 7.815  | 9.837  | 11.345 |
| 4                             | 7.779           | 9.488  | 11.668 | 13.277 |
| 5                             | 9.236           | 11.070 | 13.388 | 15.086 |
| 6                             | 10.645          | 12.592 | 15.033 | 16.812 |
| 7                             | 12.017          | 14.067 | 16.622 | 18.475 |
| 8                             | 13.362          | 15.507 | 18.168 | 20.090 |
| 9                             | 14.684          | 16.919 | 19.679 | 21.666 |
| 10                            | 15.987          | 18.307 | 21.161 | 23.209 |
| 11                            | 17.275          | 19.675 | 22.618 | 24.725 |
| 12                            | 18.549          | 21.026 | 24.054 | 26.217 |
| 13                            | 19.812          | 22.362 | 25.472 | 27.688 |
| 14                            | 21.064          | 23.685 | 26.873 | 29.141 |
| 15                            | 22.307          | 24.996 | 28.259 | 30.578 |
| 16                            | 23.542          | 26.296 | 29.633 | 32.000 |
| 17                            | 24.769          | 27.587 | 30.995 | 33.409 |
| 18                            | 25.989          | 28.869 | 32.346 | 34.805 |
| 19                            | 27.204          | 30.144 | 33.687 | 36.191 |
| 20                            | 28.412          | 31.410 | 35.020 | 37.566 |
| 21                            | 29.615          | 32.671 | 36.343 | 38.932 |
| 22                            | 30.813          | 33.924 | 37.659 | 40.289 |
| 23                            | 32.007          | 35.172 | 38.968 | 41.638 |
| 24                            | 33.196          | 36.415 | 40.270 | 42.980 |
| 25                            | 34.382          | 37.652 | 41.566 | 44.314 |
| 26                            | 35.563          | 38.885 | 42.856 | 45.642 |
| 27                            | 36.741          | 40.113 | 44.140 | 46.963 |
| 28                            | 37.916          | 41.337 | 45.419 | 48.278 |
| 29                            | 39.087          | 42.557 | 46.693 | 49.588 |
| 30                            | 40.256          | 43.773 | 47.962 | 50.892 |

# APPENDIX C: SOFTWARE COMMANDS

## CHAPTER 2

- 2-1. The Excel commands to use the PivotTable Wizard to create the frequency table, bar chart, and pie chart on page 24 are:
- Open the Applewood Auto Group data file.
  - Click on a cell somewhere in the data set, such as cell C5.
  - Click on the *Insert* menu on the toolbar. Then click *PivotTable* on the far left of the ribbon.

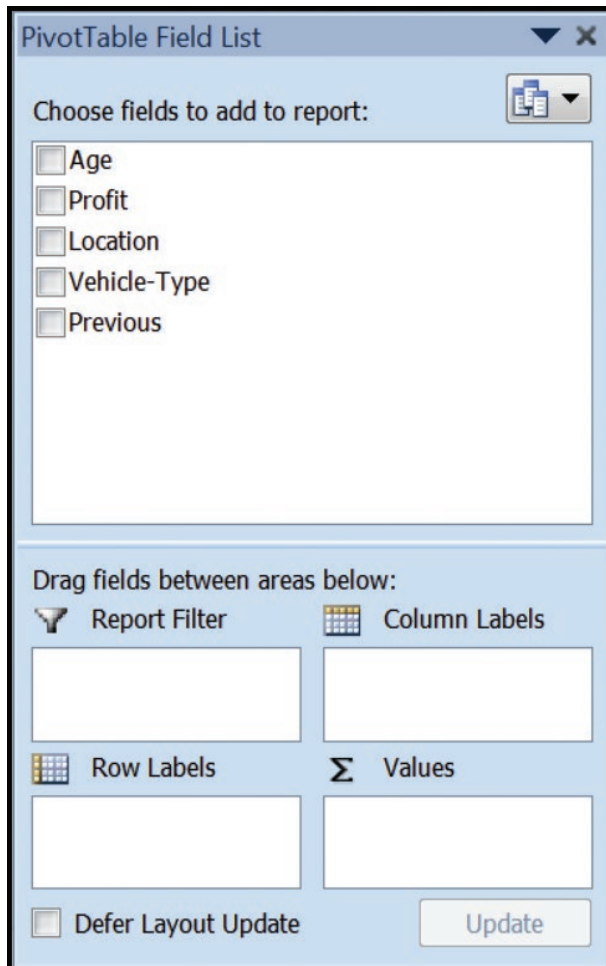
- The following screen will appear. Click on “*Select a table or range*” to select the data range as shown in the *Table/Range* row. Next, click on “*Existing Worksheet*” and select a cell location, such as *N1*, and click *OK*.



- On the right-hand side of the spreadsheet, a *PivotTable Field List* will appear with a list of the data set variables. To summarize the “Vehicle-Type” variable, click on the “Vehicle-Type” variable and it will appear in the lower left box called *Row Label*. You will note that the frequency table is started in cell *N1*, with the rows labeled with the values of the variable “Vehicle-Type.” Next, return to the top box, and select and drag the “Vehicle-Type” variable to the “ $\Sigma$  Values” box. A column of frequencies will be added to the table. Note that you can format the table to center the values and also relabel the column headings as needed.
- To create the bar chart, select any cell in the PivotTable. Next, select the *Insert* menu from the toolbar, and within the *Charts* group, select a bar chart from the *Column* drop-down menu. A bar chart appears. Click on the chart heading and label the chart as needed.
- To create the pie chart, the frequencies should be converted to relative frequencies. Click in the body of the Pivot Table and the *PivotTable Field List* will appear to the right. In the “ $\Sigma$  Values” box, click on the pull-down menu for “Count of Vehicle Type” and select the *Value Field Settings* option. You will see a number of different selections that can be used to summarize the variables in a PivotTable. Click on the tab “*Show Values As*” and, in the pull-down menu, select “% of Grand Total.” The frequencies will be converted to relative frequencies.  
To create the pie chart, select any cell in the PivotTable. Next, select the *Insert* menu from the toolbar, and within

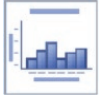
the *Charts* group, select a pie chart from the *Column* drop-down menu. A pie chart appears. Click on the chart heading and label the chart as needed. To add the percentages, click on the pie chart and a menu will appear. Click on “*Add Data Labels*.”

- 2-2. The Excel commands to use the PivotTable Wizard to create the frequency and relative frequency distributions on page 31 and the histogram on page 35 follow.
- Open the Applewood Auto Group data file.
  - Click on a cell somewhere in the data set, such as cell C5.
  - Click on the *Insert* menu on the toolbar. Then click on *PivotTable* on the far left of the ribbon.
  - The following screen will appear. Click on “*Select a table or range*” to select the data range as shown in the *Table/Range* row. Next, click on “*New Worksheet*” and the PivotTable will be created in a new worksheet.
  - On the right-hand side of the spreadsheet, a *PivotTable Field List* will appear with a list of the data set variables. To summarize the “Profit” variable, click on the “Profit” variable and drag it to the “Row Labels” box. Then return to the top box, click on “Profits” again and drag it to the “ $\Sigma$  Values” box. Staying in this box, click on the pull-down menu for “Sum of Profit.” You will see a number of different selections that can be used to summarize the variables in a PivotTable. In the “*Summarize Values As*” tab, select “Count” to create frequencies for the variable “Profit.” A PivotTable will appear in the new worksheet.



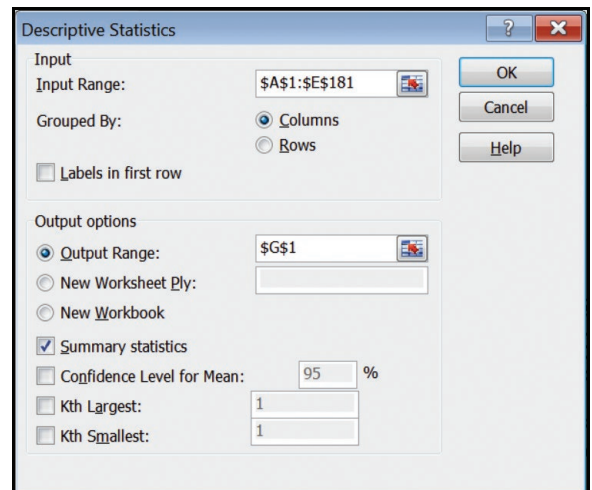
- f. In the PivotTable, the left column shows each value of the variable "Profit." To create classes for "Profit," select any cell in the column and right-click. A menu appears. Select "Group" from the menu to create the classes. First, uncheck both boxes. Then, in the dialogue box, enter the lower limit of the first class as the "Starting at" value. Enter the upper limit of the last class as the "Ending at" value. Then enter the class interval as the "By" value. Click **OK**. A frequency distribution appears.
- g. To create a relative frequency distribution, point and click on one of the cells in the PivotTable and the "PivotTable Field List" appears to the right. Click and drag the variable "Profits" to the "Σ Values" box. A second "Counts of Profit" appears. In the "Σ Values," click on the second "Counts of Profit" and select the "Value Fields Setting." You will see a number of different selections that can be used to summarize the variables in a PivotTable. Click on the tab "Show Values As" and, in the pull-down menu, select "% of Grand Total." The relative frequencies will be added to the table. You can format the table by relabeling the column headings such as "Frequency" and "Relative Frequency."
- h. To create a histogram, select a cell in the PivotTable, choose the *Insert* menu from the toolbar, and within the *Charts* group, select a *column* chart from the *Column* drop-down

menu. A histogram appears with both "Count of Profit" and "Count of Profit2." On the "Count of Profit2" bubble at the top of the chart, right-click and select "Remove Field." Then the chart and PivotTable only report the frequencies. To eliminate the space between the bars, select the entire chart area, and "PivotChart Tools" will appear at the top. Select "Design." In the "Chart Layouts" choices, select the option that shows no spaces between the bars. The option is illustrated in the figure to the right. To add data labels, select the histogram, right-click, and select "Add Data Labels." Relabel the chart and axes as needed.



### CHAPTER 3

- 3–1. The Excel Commands for the descriptive statistics on page 66 are:
- From **Connect**, retrieve the Applewood data.
  - From the menu bar, select **Data** and then **Data Analysis**. Select **Descriptive Statistics** and then click **OK**.
  - For the **Input Range**, type *C1:C181*, indicate that the data are grouped by column and that the labels are in the first row. Click on **Output Range**, indicate that the output should go in *G1* (or any place you wish), click on **Summary statistics**, and then click **OK**.

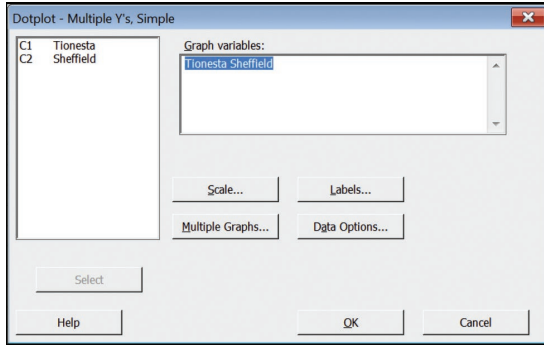


- After you get your results, double-check the count in the output to be sure it contains the correct number of items.

### CHAPTER 4

- 4–1. The Minitab commands for the dot plot on page 90 are:
- Enter the number of vehicles serviced at Tionesta Ford Lincoln in column *C1* and Sheffield Motors in *C2*. Name the variables accordingly.
  - Select **Graph** and **Dotplot**. In the first dialog box, select **Multiple Y's, Simple** in the lower left corner, and click **OK**. In the next dialog box, select **Tionesta** and **Sheffield** as the variables to **Graph**, click on **Labels**, and write an appropriate title. Then click **OK**.
  - To calculate the descriptive statistics shown in the output, select **Stat, Basic statistics**, and then **Display Descriptive statistics**. In the dialog box, select **Tionesta** and **Sheffield** as the variables, click on **Statistics**, select the desired statistics to be output, and finally click **OK** twice.





4-2. The Minitab commands for the descriptive summary on page 94 are:

- a. Input the data on the Morgan Stanley commissions from the Example on page 93.
- b. From the toolbar, select **Stat, Basic Statistics**, and **Display Descriptive Statistics**. In the dialog box, select **Commissions** as the **Variable**, and then click **OK**.

4-3. The Excel commands for the quartiles on page 95 follow.

a. Input the data on the Morgan Stanley commissions from the example on page 94 in column A.

If you are using Excel 2010 and wish to compute quartiles using formula (4-1), the steps are:

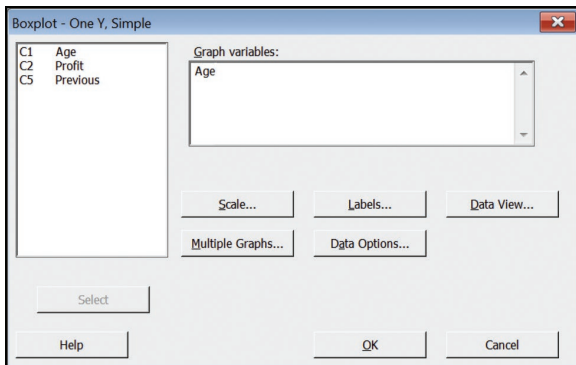
- b. In cell C3 type **Formula (4-1)**, in C4 type **Quartile 1**, and in C6 type **Quartile 3**.
- c. In cell D4 type “=QUARTILE.EXC(A2:A16,1)” and hit **Enter**. In cell D6 type “=QUARTILE.EXC(A2:A16,3)” and hit **Enter**.

If you are using either Excel 2007 or 2010 and wish to compute quartiles using the Excel Method:

- b. In cell C8 type **Excel Method**, in C9 type **Quartile 1**, and in C11 type **Quartile 3**.
- c. In cell D8 type “=QUARTILE(A2:A16,1)” and hit **Enter**. In cell D11 type “=QUARTILE(A2:A16,3)” and hit **Enter**.

4-4. The Minitab commands for the box plot on page 98 are:

- a. Import the Applewood Auto Group data from Connect.



- b. Select **Graph** and then **Boxplot**. In the dialog box, select **Simple** in the upper left corner and click **OK**. Select **Age** as the **Graph Variable**, click on **Labels** and include an appropriate heading, and then click **OK**.

4-5. The Minitab commands for the descriptive summary on page 103 are:

- a. Enter the data in the first column. In the cell below C1, enter the variable *Earnings*.

- b. Select **Stat, Basic Statistics**, and then click on **Graphical Summary**. Select **Earnings** as the variable, and then click **OK**.

4-6. The Excel commands for the scatter diagram on page 106 are:

- a. Retrieve the Applewood Auto Group data.
- b. Using the mouse, highlight the column of age and profit. Include the first row.
- c. Select the **Insert** tab. Select **Scatter** from the **Chart** options. Select the top left option. The scatter plot will appear.
- d. With **Chart Tools** displayed at the top, select the **Layout** tab. Select **Chart Title** and type in a title for the plot. Next, under the same **Layout** tab, select **Axis Titles**. Using **Primary Vertical Axis Title**, name the vertical axis *Profit*. Using the **Primary Horizontal Axis Title**, name the horizontal axis *Age*. Next, select **Legend** and select **None**.

## CHAPTER 5

5-1. The Excel Commands to determine the number of permutations shown on page 146 are:

- a. Click on the **Formulas** tab in the top menu, then, on the far left, select **Insert Function fx**.
- b. In the **Insert Function** box, select **Statistical** as the category, then scroll down to **PERMUT** in the **Select a function** list. Click **OK**.
- c. In the **PERM** box after **Number**, enter 8, and in the **Number\_chosen** box, enter 3. The correct answer of 336 appears twice in the box.

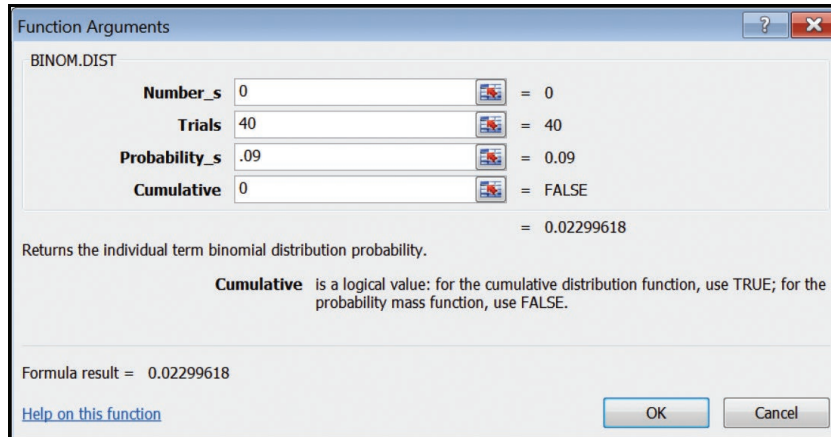
5-2. The Excel commands to determine the number of combinations shown on page 146 are:

- a. Click on the **Formulas** tab in the top menu, then, on the far left, select **Insert Function fx**.
- b. In the **Insert Function** box, select **Math & Trig** as the category, then scroll down to **COMBIN** in the **Select a function** list. Click **OK**.
- c. In the **COMBIN** box after **Number**, enter 7, and in the **Number\_chosen** box, enter 3. The correct answer of 35 appears twice in the box.

## CHAPTER 6

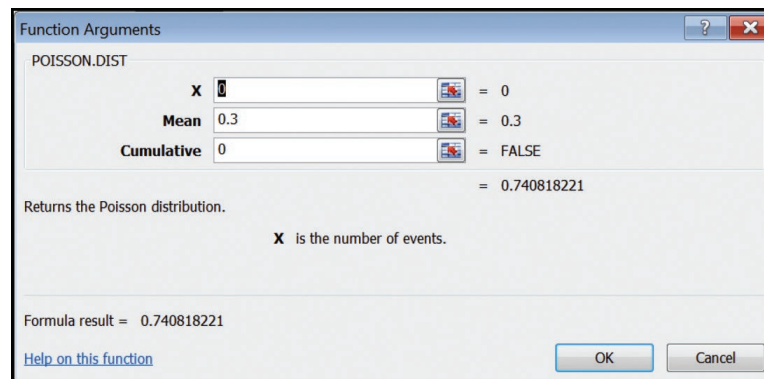
6-1. The Excel commands necessary to determine the binomial probability distribution on page 168 are:

- a. On a blank Excel worksheet, write the word *Success* in cell A1 and the word *Probability* in B1. In cells A2 through A17, write the integers 0 to 15. Click on B2 as the active cell.
- b. Click on the **Formulas** tab in the top menu, then, on the far left, select **Insert Function fx**.
- c. In the first dialog box, select **Statistical** in the function category and **BINOM.DIST** in the function name category, then click **OK**.
- d. In the second dialog box, enter the four items necessary to compute a binomial probability.
  1. Enter 0 for the **Number\_s**.
  2. Enter 40 for the **Trials**.
  3. Enter .09 for the probability of a success.
  4. Enter the word *false* or the number 0 for **Cumulative** and click on **OK**.
  5. Excel will compute the probability of 0 successes in 40 trials, with a .09 probability of success. The result, .02299618, is stored in cell B2.
- e. To complete the probability distribution for successes of 1 through 15, double-click on cell B2. The binomial function should appear. Replace the 0 to the right of the open parentheses with the cell reference A2.
- f. Move the mouse to the lower right corner of cell B2 until a solid black + symbol appears, then click and hold and highlight the B column to cell B17. The probability of a success for the various values of the random variable will appear.



- 6–2. The Excel commands necessary to determine the Poisson probability distribution on page 175 are:
- On a blank Excel worksheet, write the word *Success* in cell **A1** and the word *Probability* in **B1**. In cells **A2** through **A9**, write the integers 0 to 7. Click on **B2** as the active cell.
  - Click on the **Formulas** tab in the top menu, then, on the far left, select **Insert Function fx**.
  - In the first dialog box, select **Statistical** in the function category and **POISSON.DIST** in the function name category, then click **OK**.
  - In the second dialog box, enter the three items necessary to compute a Poisson probability.
    - Enter 0 for **X**.
    - Enter 0.3 for the **Mean**.

- Enter the word *false* or the number 0 for **Cumulative** and click **OK**.
- Excel will compute the probability of 0 successes for a Poisson probability distribution with a mean of 0.3. The result, .74081822, is stored in cell **B2**.
- To complete the probability distribution for successes of 1 through 7, double-click on cell **B2**. The Poisson function should appear. Replace the 0 to the right of the open parentheses with the cell reference **A2**.
- Move the mouse to the lower right corner of cell **B2** until a solid black + symbol appears, then click and hold and highlight the **B** column to cell **B9**. The probability of a success for the various values of the random variable will appear.



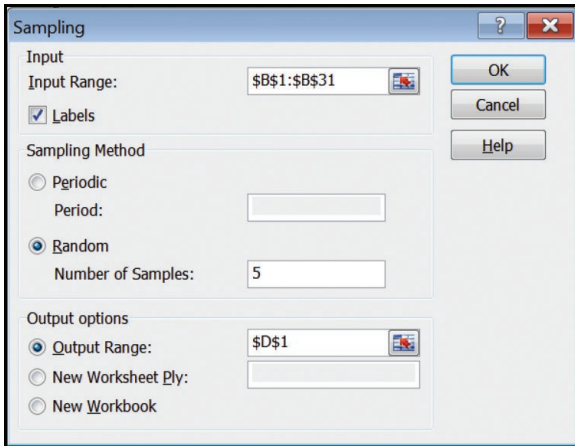
## CHAPTER 7

- 7–1. The Excel commands necessary to produce the output on page 197 are:
- Click on the **Formulas** tab in the top menu, then, on the far left, select **Insert Function fx**. Then from the category box, select **Statistical** and below that, **NORM.DIST**, and click **OK**.
  - In the dialog box, put 1100 in the box for **X**, 1000 for the **Mean**, 100 for the **Standard\_dev**, and *True* in the **Cumulative** box, and click **OK**.
  - The result will appear in the dialog box. If you click **OK**, the answer appears in your spreadsheet.

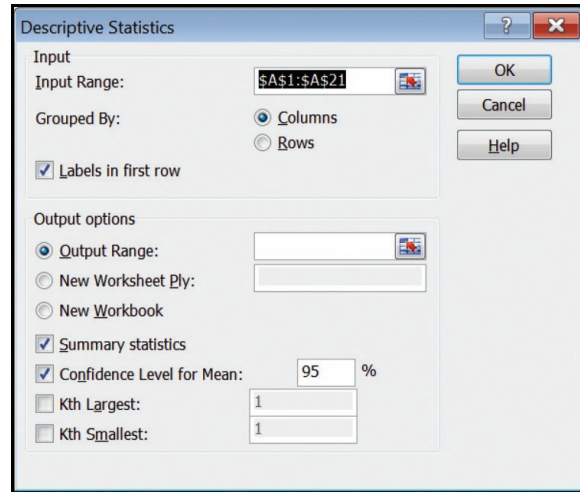
- 7–2. The Excel commands necessary to produce the output on page 203 are:
- Click the **Formulas** tab in the top menu, then, on the far left, select **Insert Function fx**. Then from the category box, select **Statistical** and below that, **NORM.INV**, and click **OK**.
  - In the dialog box, set the **Probability** to .04, the **Mean** to 67900, and the **Standard\_dev** to 2050.
  - The results will appear in the dialog box. Note that the answer is different from page 203 because of rounding error. If you click **OK**, the answer also appears in your spreadsheet.
  - Try entering a **Probability** of .04, a **Mean** of 0, and a **Standard\_dev** of 1. The z-value will be computed.

## CHAPTER 8

- 8-1. The Excel commands to select a simple random sample from the rental data on page 214 are:
- Select the **Data** tab on the top of the menu. Then on the far right select **Data Analysis**, then **Sampling** and **OK**.
  - For **Input Range**, insert **B1:B31**. Since the column is named, click the **Labels** box. Select **Random**, and enter the sample size for the **Number of Samples**, in this case 5. Click on **Output Range** and indicate the place in the spreadsheet where you want the sample information. Note that your sample results will differ from those in the text. Also recall that Excel samples with replacement, so it is possible for a population value to appear more than once in the sample.

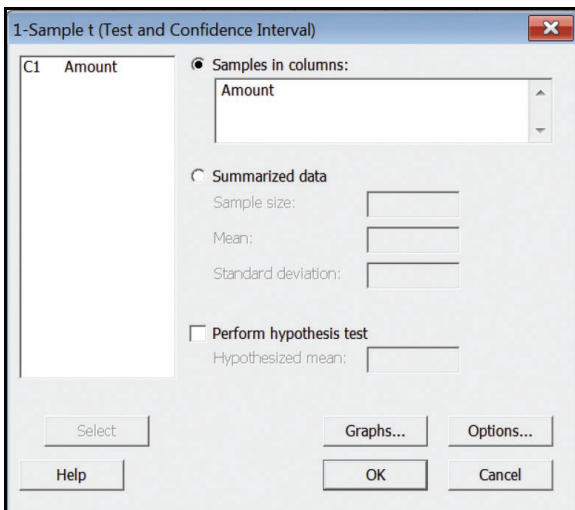


- 9-2. The Excel commands for the confidence interval for the amounts spent at the Inlet Square Mall on page 258 are:
- Select the **Data** tab on the top menu. Then, on the far right, select **Data Analysis**, and then **Descriptive Statistics**, and click **OK**.
  - For the **Input Range**, type **A1:A21**, click on **Labels in first row**, type **C1** as the **Output Range**, click on **Summary statistics** and **Confidence Level for Mean**, and then click on **OK**.



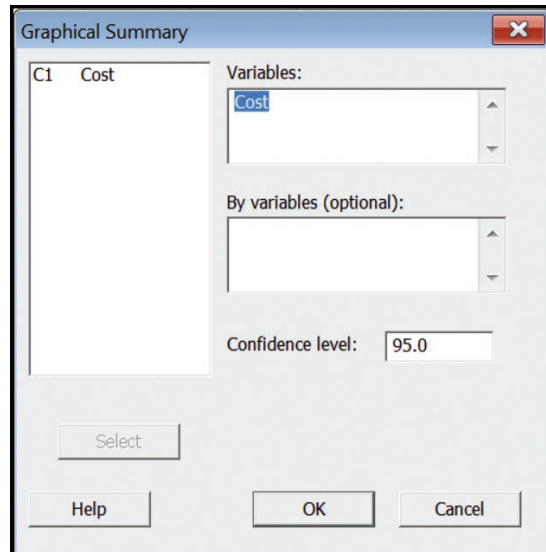
## CHAPTER 9

- 9-1. The Minitab commands for the confidence interval for the amount spent at the Inlet Square Mall on page 257 are:
- Enter the 20 amounts spent in column **C1** and name the variable **Amount**.
  - On the toolbar, select **Stat, Basic Statistics**, and click on **1-Sample t**.
  - Select **Samples in columns:** and select **Amount** and click **OK**.

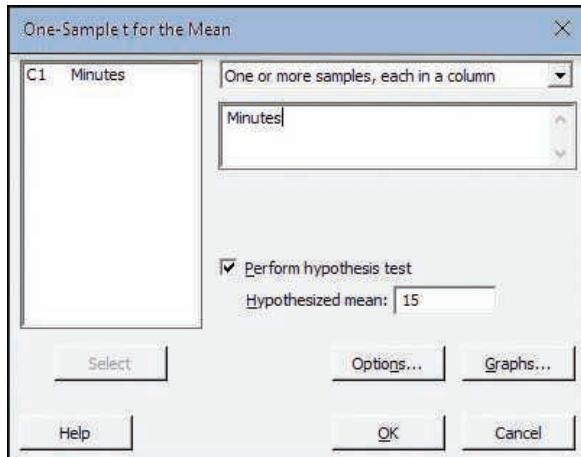


## CHAPTER 10

- 10-1. The Minitab commands for the histogram and the descriptive statistics on page 291 are:
- Enter the 26 sample observations in column **C1** and name the variable **Cost**.
  - From the menu bar, select **Stat, Basic Statistics**, and **Graphical Summary**. In the dialog box, select **Cost** as the variable and click **OK**.

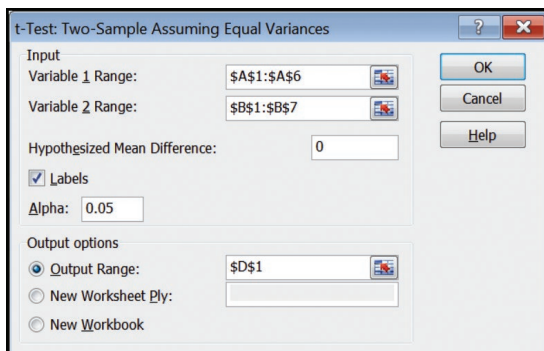


- 10–2. The Minitab commands for the one-sample *t* test on page 296 are:
- Enter the sample data into column **C1** and name the variable *Minutes*.
  - From the menu bar, select **Stat, Basic Statistics**, and **1-Sample *t***, and then hit **Enter**.
  - Select **Minutes** as the variable, select **Perform hypothesis test**, and insert the value **15**. Click **Options**. Under **Alternate**, select **greater than**. Finally, click **OK** twice.

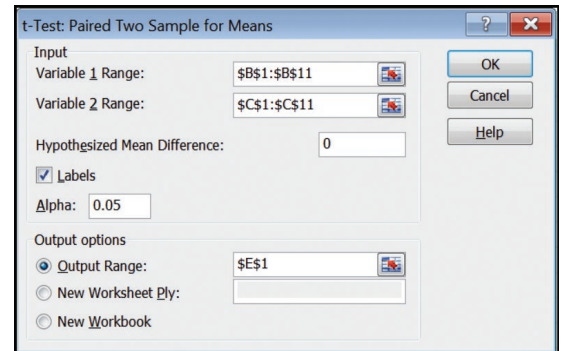


## CHAPTER 11

- 11–1. The Excel commands for the two-sample *t*-test on page 316 are:
- Enter the data into columns A and B (or any other columns) in the spreadsheet. Use the first row of each column to enter the variable name.
  - Select the **Data** tab on the top menu. Then, on the far right, select **Data Analysis**. Select **t-Test: Two Sample Assuming Equal Variances**, and then click **OK**.
  - In the dialog box, indicate that the range of **Variable 1** is from **A1** to **A6** and **Variable 2** from **B1** to **B7**, the **Hypothesized Mean Difference** is **0**, click **Labels**, **Alpha** is **0.05**, and the **Output Range** is **D1**. Click **OK**.

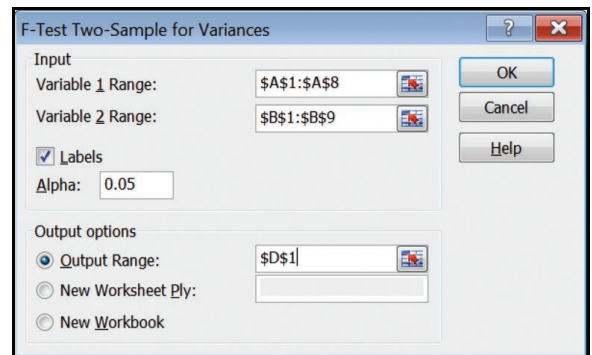


- 11–2. The Excel commands for the paired *t*-test on page 321 are:
- Enter the data into columns B and C (or any other two columns) in the spreadsheet, with the variable names in the first row.
  - Select the **Data** tab on the top menu. Then, on the far right, select **Data Analysis**. Select **t-Test: Paired Two Sample for Means**, and then click **OK**.
  - In the dialog box, indicate that the range of **Variable 1** is from **B1** to **B11** and **Variable 2** from **C1** to **C11**, the **Hypothesized Mean Difference** is **0**, click **Labels**, **Alpha** is **.05**, and the **Output Range** is **E1**. Click **OK**.

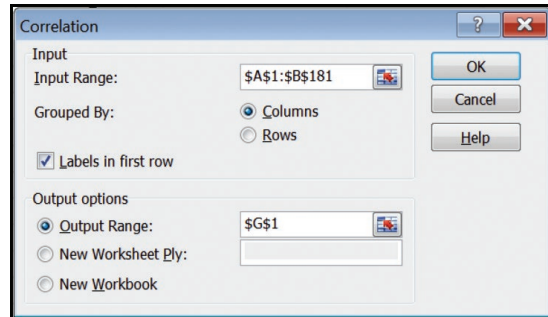
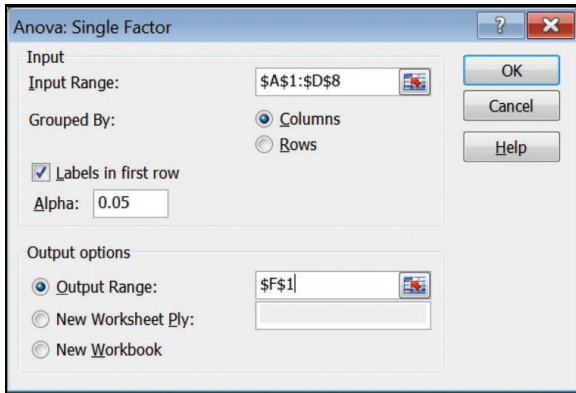


## CHAPTER 12

- 12–1. The Excel commands for the test of variances on page 339 are:
- Enter the data for U.S. 25 in column A and for I-75 in column B. Label the two columns.
  - Select the **Data** tab on the top menu. Then, on the far right, select **Data Analysis**. Select **F-Test: Two-Sample for Variances**, then click **OK**.
  - The range of the first variable is **A1:A8**, and **B1:B9** for the second. Click on **Labels**, enter **0.05** for **Alpha**, select **D1** for the **Output Range**, and click **OK**.



- 12–2. The Excel Commands for the one-way ANOVA on page 348 are:
- Input the data into four columns labeled *Northern*, *WTA*, *Pocono*, and *Branson*.
  - Select the **Data** tab on the top menu. Then, on the far right, select **Data Analysis**. Select **ANOVA: Single Factor**, then click **OK**.
  - In the subsequent dialog box, make the input range **A1:D8**, click on **Grouped By: Columns**, click on **Labels in first row**, enter **0.05** in Alpha, and finally select **Output Range** as **F1** and click **OK**.

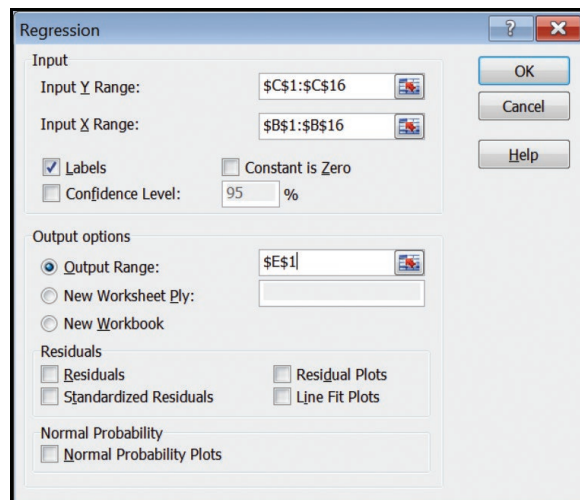
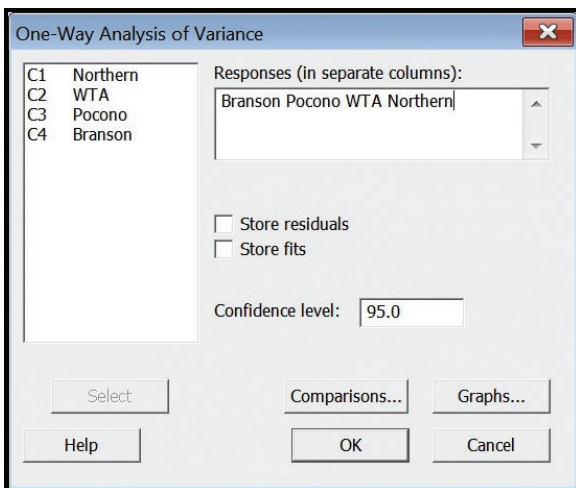


12–3. The Minitab commands for the pairwise comparisons on page 351 are:

- Input the data into four columns and identify the columns as *Northern*, *WTA*, *Pocono*, and *Branson*.
- Select **Stat**, **ANOVA**, and **One-way**, select “Response data are in a separate column for each factor level,” select and enter the variable names into the **Responses** box by clicking on the variable names in the following order: *Branson*, *Pocono*, *WTA*, and *Northern*. Then select **Comparisons** and then select **Fisher’s**, **individual error rate**. Then click **OK**.

13–2. The computer commands for the Excel output on page 389 are:

- Enter the variable names in row 1 of columns A, B, and C. Enter the data in rows 2 through 16 in the same columns.
- Select the **Data** tab on the top of the menu. Then, on the far right, select **Data Analysis**. Select **Regression**, then click **OK**.
- For our spreadsheet, we have *Calls* in column B and *Sales* in column C. The **Input Y Range** is *C1:C16* and the **Input X Range** is *B1:B16*. Click on **Labels**, select *E1* as the **Output Range**, and click **OK**.



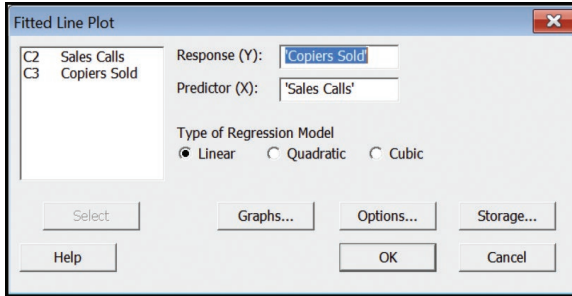
## CHAPTER 13

13–1. The Excel commands for calculating the correlation coefficient on page 374 are:

- Access the Applewood Auto Group data in Connect.
- Select the **Data** tab on the top of the ribbon. Then, on the far right, select **Data Analysis**. Select **Correlation**, and click **OK**.
- For the **Input Range**, highlight the **Age** and **Profit** columns, including the labels in row 1. The data are grouped by **Columns**. Check the **Labels in first row** box. Select a cell in the worksheet as the beginning of the range to output the correlation. Click **OK**.

13–3. The Minitab commands for the confidence intervals and prediction intervals on page 399 are:

- Select **Stat**, **Regression**, and **Fitted line plot**.
- In the next dialog box, the **Response (Y)** is *Copiers Sold* and **Predictor (X)** is *Sales Calls*. Select **Linear** for the type of regression model and then click on **Options**.
- In the **Options** dialog box, click on **Display confidence interval and prediction interval**, enter *95.0* in **Confidence Level**, type an appropriate heading in the **Title** box, then click **OK** and then **OK** again.

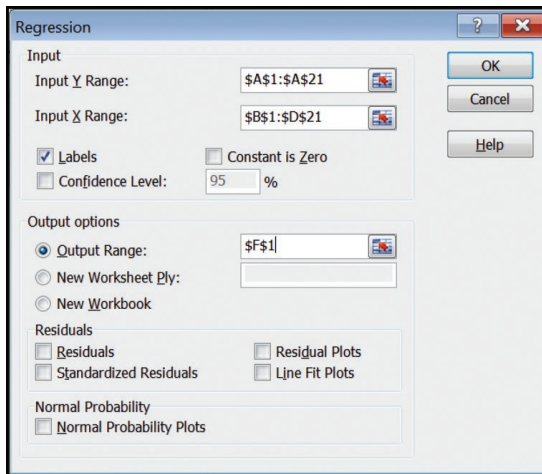


## CHAPTER 14

Note: We do not show steps for all the statistical software in Chapter 14. The following shows the basic steps.

14–1. The Excel commands to produce the multiple regression output on page 422 are:

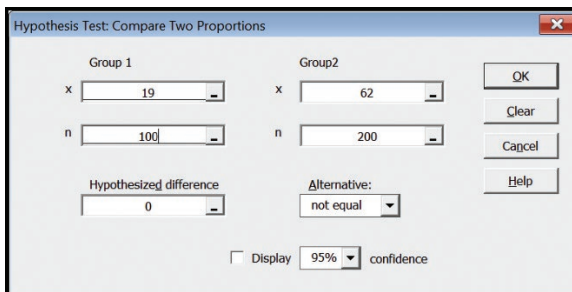
- Import the data from Connect. The file name is **Tbl14**.
- Select the **Data** tab on the top menu. Then on the far right, select **Data Analysis**. Select **Regression** and click **OK**.
- Make the **Input Y Range** **A1:A21**, the **Input X Range** **B1:D21**, check the **Labels** box, the **Output Range** is **F1**, then click **OK**.



## CHAPTER 15

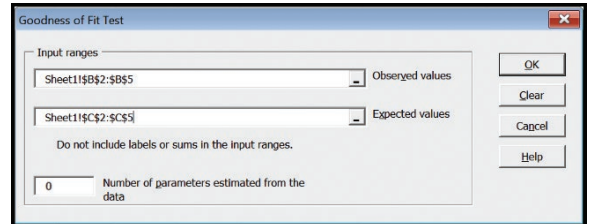
15–1. The MegaStat commands for the two-sample test of proportions on page 477 are:

- Select **MegaStat** from the **Add-Ins** tab. From the menu, select **Hypothesis Tests**, and then **Compare Two Independent Proportions**.
- Enter the data. For **Group 1**, enter **x** as **19** and **n** as **100**. For **Group 2**, enter **x** as **62** and **n** as **200**. Select **OK**.



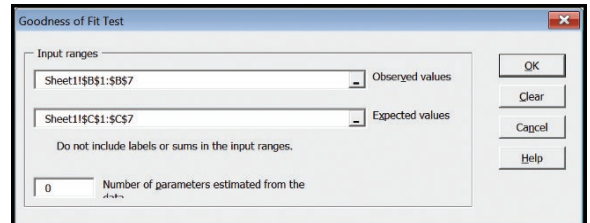
15–2. The MegaStat commands to create the chi-square goodness-of-fit test on page 483 are:

- Enter the information from Table 15–2 into a worksheet as shown.
- Select **MegaStat**, **Chi-Square/Crosstabs**, and **Goodness of Fit Test**, and hit **Enter**.
- In the dialog box, select **B2:B5** as the **Observed values**, **C2:C5** as the **Expected values**, and enter **0** as the **Number of parameters estimated from the data**. Click **OK**.



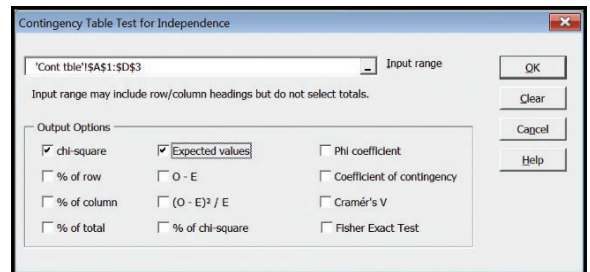
15–3. The MegaStat commands to create the chi-square goodness-of-fit tests on pages 488 and 489 are the same except for the number of items in the observed and expected frequency columns. Only one dialog box is shown.

- Enter the Levels of Management information shown on page 488.
- Select **MegaStat**, **Chi-Square/Crosstabs**, and **Goodness of Fit Test**, and hit **Enter**.
- In the dialog box, select **B1:B7** as the **Observed values**, **C1:C7** as the **Expected values**, and enter **0** as the **Number of parameters estimated from the data**. Click **OK**.



15–4. The MegaStat commands for the contingency table analysis on page 493 are:

- Enter Table 15–5 on page 491 into cells **A1** through **D3**. Include the row and column labels. DO NOT include the Total column or row.
- Select **MegaStat** from the **Add-Ins** tab. From the menu, select **Chi-square/Crosstab**, then select **Contingency Table**.
- For the **Input range**, select cells **A1** through **D3**. Check the **chi-square** and **Expected values** boxes. Select **OK**.



# APPENDIX D: ANSWERS TO ODD-NUMBERED CHAPTER EXERCISES & SOLUTIONS TO PRACTICE TESTS

## Answers to Odd-Numbered Chapter Exercises

### CHAPTER 1

1. a. Interval                      d. Nominal  
b. Ratio                          e. Ordinal  
c. Nominal                      f. Ratio
3. Answers will vary.
5. Qualitative data are not numerical, whereas quantitative data are numerical. Examples will vary by student.
7. A discrete variable may assume only certain values. A continuous variable may assume an infinite number of values within a given range. The number of traffic citations issued each day during February in Garden City Beach, South Carolina, is a discrete variable. The weight of commercial trucks passing the weigh station at milepost 195 on Interstate 95 in North Carolina is a continuous variable.
9. a. Ordinal  
b. Ratio  
c. The newer system provides information on the distance between exits.
11. If you were using this store as typical of all Best Buy stores, then the daily number sold last month would be a sample. However, if you considered the store as the only store of interest, then the daily number sold last month would be a population.
- 13.

|              | Discrete Variable   | Continuous Variable         |
|--------------|---|-----------------------------|
| Qualitative  | b. Gender<br>d. Soft drink preference<br>g. Student rank in class<br>h. Rating of a finance professor |                             |
| Quantitative | c. Sales volume of MP3 players<br>f. SAT scores<br>i. Number of home video screens                    | a. Salary<br>e. Temperature |

|          | Discrete Variable  | Continuous Variable |
|----------|--|---------------------|
| Nominal  | b. Gender  |                     |
| Ordinal  | d. Soft drink preference<br>g. Student rank in class<br>h. Rating of a finance professor |                     |
| Interval | f. SAT scores  | e. Temperature      |
| Ratio    | c. Sales volume of MP3 players<br>i. Number of home video screens                        | a. Salary           |

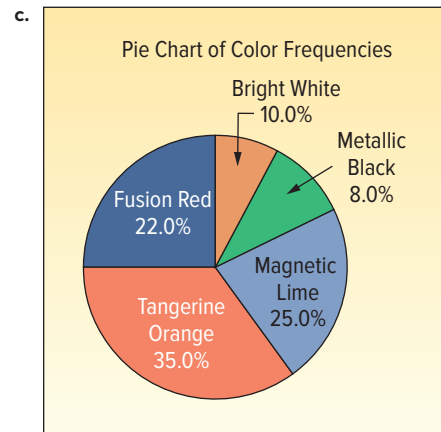
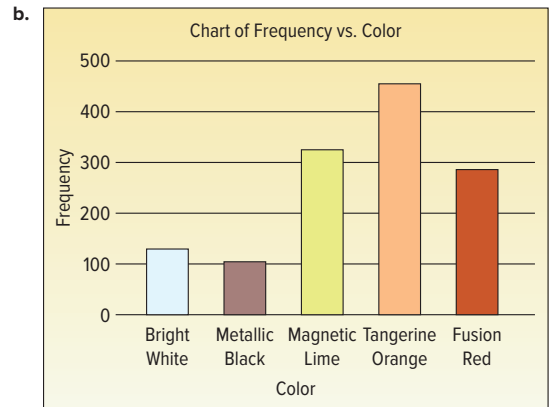
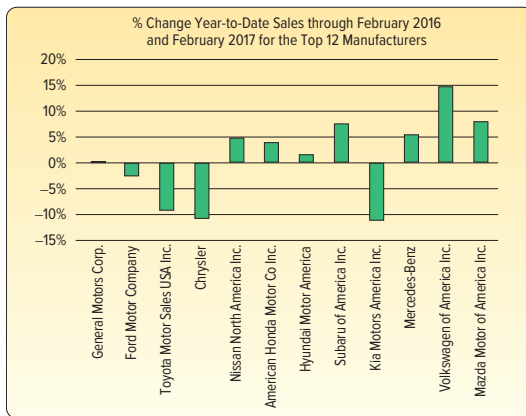
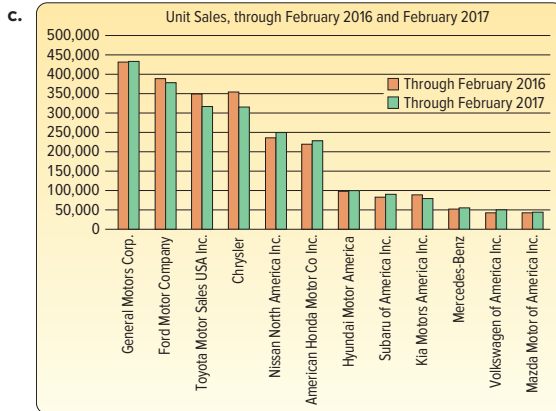
15. According to the sample information, 120/300, or 40%, would accept a job transfer.
17. a.

| Manufacturer                 | Difference |
|------------------------------|------------|
| Nissan North America Inc.    | 11,414     |
| American Honda Motor Co Inc. | 8,584      |
| Subaru of America Inc.       | 6,267      |
| Volkswagen of America Inc.   | 6,255      |
| Jaguar                       | 3,559      |
| Audi of America Inc.         | 3,374      |
| Mazda Motor of America Inc.  | 3,275      |
| Mitsubishi Motors N A Inc.   | 3,247      |

| Manufacturer                | Difference |
|-----------------------------|------------|
| Tesla                       | 2,925      |
| Mercedes-Benz               | 2,838      |
| Hyundai Motor America       | 1,507      |
| General Motors Corp.        | 1,479      |
| Maserati                    | 723        |
| Alfa Romeo                  | 436        |
| Porsche Cars NA Inc.        | 324        |
| Bentley                     | 160        |
| Ferrari                     | 159        |
| Rolls Royce                 | 98         |
| BMW of North America Inc.   | 87         |
| Lamborghini                 | -28        |
| Smart                       | -149       |
| Land Rover                  | -162       |
| Fiat                        | -721       |
| Mini                        | -813       |
| Volvo                       | -1,381     |
| Kia Motors America Inc.     | -9,743     |
| Ford Motor Company          | -9,873     |
| Toyota Motor Sales USA Inc. | -31,851    |
| Chrysler                    | -37,841    |

- b. Percentage differences with top five and bottom five.

| Manufacturer                 | % Change from 2016 to 2017 |
|------------------------------|----------------------------|
| Alfa Romeo                   | 379%                       |
| Jaguar                       | 124                        |
| Bentley                      | 112                        |
| Rolls Royce                  | 67                         |
| Tesla                        | 61                         |
| Maserati                     | 58                         |
| Ferrari                      | 51                         |
| Mitsubishi Motors N A Inc.   | 23                         |
| Volkswagen of America Inc.   | 15                         |
| Audi of America Inc.         | 14                         |
| Mazda Motor of America Inc.  | 8                          |
| Subaru of America Inc.       | 8                          |
| Mercedes-Benz                | 5                          |
| Nissan North America Inc.    | 5                          |
| Porsche Cars NA Inc.         | 4                          |
| American Honda Motor Co Inc. | 4                          |
| Hyundai Motor America        | 2                          |
| General Motors Corp.         | 0                          |
| BMW of North America Inc.    | 0                          |
| Land Rover                   | -1                         |
| Ford Motor Company           | -3                         |
| Toyota Motor Sales USA Inc.  | -9                         |
| Chrysler                     | -11                        |
| Kia Motors America Inc.      | -11                        |
| Mini                         | -13                        |
| Fiat                         | -14                        |
| Volvo                        | -15                        |
| Smart                        | -18                        |
| Lamborghini                  | -19                        |



19. Earnings varied over the last 12 years. In 2008, there was a very large increase followed by a large decrease in 2009. Perhaps the earnings were affected by the financial “collapse” during the years 2008–2010. Over the last seven years, there was first an increasing trend in earnings followed by a decrease in earnings. Could these trends be explained by the price of oil?
21. a. League is a qualitative variable; the others are quantitative.  
b. League is a nominal-level variable; the others are ratio-level variables.

## CHAPTER 2

1. 25% market share.  
3.

| Season | Frequency | Relative Frequency |
|--------|-----------|--------------------|
| Winter | 100       | .10                |
| Spring | 300       | .30                |
| Summer | 400       | .40                |
| Fall   | 200       | .20                |
| Total  | 1,000     | 1.00               |

5. a. A frequency table.

| Color            | Frequency | Relative Frequency |
|------------------|-----------|--------------------|
| Bright White     | 130       | 0.10               |
| Metallic Black   | 104       | 0.08               |
| Magnetic Lime    | 325       | 0.25               |
| Tangerine Orange | 455       | 0.35               |
| Fusion Red       | 286       | 0.22               |
| Total            | 1,300     | 1.00               |

- d. 350,000 orange, 250,000 lime, 220,000 red, 100,000 white, and 80,000 black, found by multiplying relative frequency by 1,000,000 production.
7.  $2^5 = 32$ ,  $2^6 = 64$ , therefore, 6 classes
9.  $2^7 = 128$ ,  $2^8 = 256$ , suggests 8 classes  
 $i \geq \frac{\$567 - \$235}{8} = 41$  Class intervals of 45 or 50 would be acceptable.

11. a.  $2^4 = 16$ , suggests 5 classes  
b.  $i \geq \frac{31 - 25}{5} = 1.2$  Use interval of 1.5.

c. 24

d.

| Units           | <i>f</i> | Relative Frequency |
|-----------------|----------|--------------------|
| 24.0 up to 25.5 | 2        | 0.125              |
| 25.5 up to 27.0 | 4        | 0.250              |
| 27.0 up to 28.5 | 8        | 0.500              |
| 28.5 up to 30.0 | 0        | 0.000              |
| 30.0 up to 31.5 | 2        | 0.125              |
| Total           | 16       | 1.000              |

- e. The largest concentration is in the 27.0 up to 28.5 class (8).

13. a.

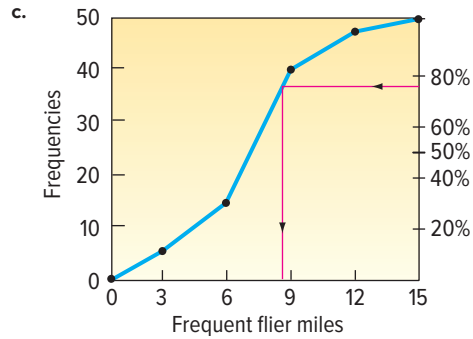
| Number of Visits | <i>f</i> |
|------------------|----------|
| 0 up to 3        | 9        |
| 3 up to 6        | 21       |
| 6 up to 9        | 13       |
| 9 up to 12       | 4        |
| 12 up to 15      | 3        |
| 15 up to 18      | 1        |
| Total            | 51       |



- b. The largest group of shoppers (21) shop at the BiLo Supermarket 3, 4, or 5 times during a one-month period. Some customers visit the store only 1 time during the month, but others shop as many as 15 times.

c.

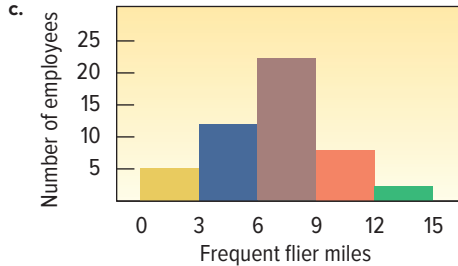
| Number of Visits | Percent of Total |
|------------------|------------------|
| 0 up to 3        | 17.65            |
| 3 up to 6        | 41.18            |
| 6 up to 9        | 25.49            |
| 9 up to 12       | 7.84             |
| 12 up to 15      | 5.88             |
| 15 up to 18      | 1.96             |
| Total            | 100.00           |



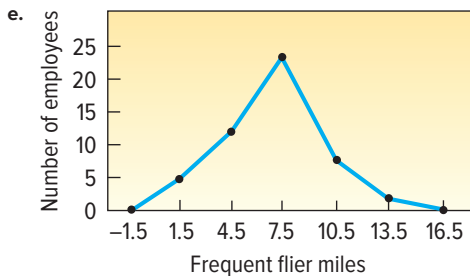
- d. About 8.7 thousand miles
23. a. A qualitative variable uses either the nominal or ordinal scale of measurement. It is usually the result of counts. Quantitative variables are either discrete or continuous. There is a natural order to the results for a quantitative variable. Quantitative variables can use either the interval or ratio scale of measurement.
- b. Both types of variables can be used for samples and populations.

15. a. Histogram  
b. 100  
c. 5  
d. 28  
e. 0.28  
f. 12.5  
g. 13

17. a. 50  
b. 1.5 thousand miles, or 1,500 miles.



- d.  $X = 1.5, Y = 5$



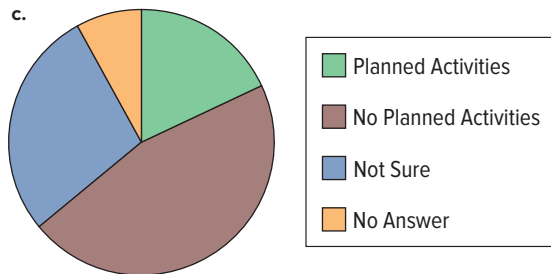
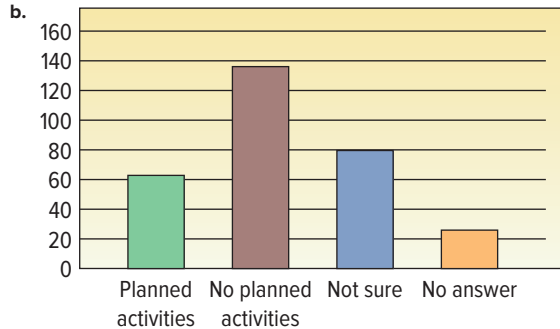
- f. For the 50 employees, about half traveled between 6,000 and 9,000 miles. Five employees traveled less than 3,000 miles, and two traveled more than 12,000 miles.

19. a. 40  
b. 5  
c. 11 or 12  
d. About \$18/hr  
e. About \$9/hr  
f. About 75%

21. a. 5  
b.

| Miles        | CF |
|--------------|----|
| Less than 3  | 5  |
| Less than 6  | 17 |
| Less than 9  | 40 |
| Less than 12 | 48 |
| Less than 15 | 50 |

25. a. Frequency table



- d. A pie chart would be better because it clearly shows that nearly half of the customers prefer no planned activities.

27.  $2^6 = 64$  and  $2^7 = 128$ , suggest 7 classes

29. a. 5, because  $2^4 = 16 < 25$  and  $2^5 = 32 > 25$

b.  $i \geq \frac{48 - 16}{5} = 6.4$  Use interval of 7.

- c. 15

d.

| Class       | Frequency  |
|-------------|------------|
| 15 up to 22 | III 3      |
| 22 up to 29 | IIII III 8 |
| 29 up to 36 | IIII II 7  |
| 36 up to 43 | IIII 5     |
| 43 up to 50 | II 2       |
|             | <hr/> 25   |

- e. It is fairly symmetric, with most of the values between 22 and 36.

31. a.  $2^5 = 32$ ,  $2^6 = 64$ , 6 classes recommended

b.  $i = \frac{10 - 1}{6} = 1.5$  Use an interval of 2.

c. 0

d.

| Class       | Frequency |
|-------------|-----------|
| 0 up to 2   | 1         |
| 2 up to 4   | 5         |
| 4 up to 6   | 12        |
| 6 up to 8   | 17        |
| 8 up to 10  | 8         |
| 10 up to 12 | 2         |

e. The distribution is fairly symmetric or bell-shaped with a large peak in the middle of the two classes of 4 up to 8.

33.

| Class           | Frequency |
|-----------------|-----------|
| 0 up to 200     | 19        |
| 200 up to 400   | 1         |
| 400 up to 600   | 4         |
| 600 up to 800   | 1         |
| 800 up to 1,000 | 2         |

This distribution is positively skewed with a large “tail” to the right or positive values. Notice that the top 7 tunes account for 4,342 plays out of a total of 5,968, or about 73% of all plays.

35. a. 56

b. 10 (found by  $60 - 50$ )

c. 55

d. 17

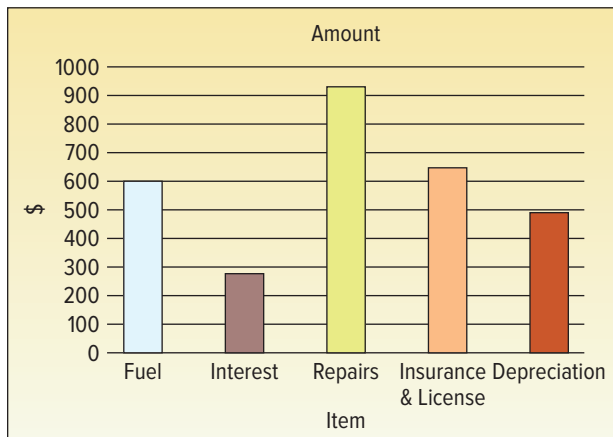
37. a. Use \$35 because the minimum is  $(\$265 - \$82)/6 = \$30.5$ .

b.

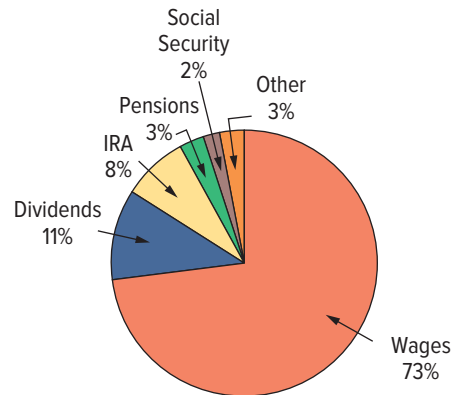
|                   |    |
|-------------------|----|
| \$ 70 up to \$105 | 4  |
| 105 up to 140     | 17 |
| 140 up to 175     | 14 |
| 175 up to 210     | 2  |
| 210 up to 245     | 6  |
| 245 up to 280     | 1  |

c. The purchases range from a low of about \$70 to a high of about \$280. The concentration is in the \$105 up to \$140 and \$140 up to \$175 classes.

39. Bar charts are preferred when the goal is to compare the actual amount in each category.



41.

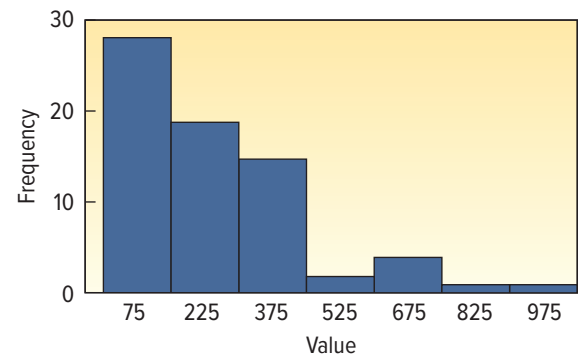


| SC Income       | Percent | Cumulative |
|-----------------|---------|------------|
| Wages           | 73      | 73         |
| Dividends       | 11      | 84         |
| IRA             | 8       | 92         |
| Pensions        | 3       | 95         |
| Social Security | 2       | 97         |
| Other           | 3       | 100        |

By far the largest part of income in South Carolina is wages. Almost three-fourths of the adjusted gross income comes from wages. Dividends and IRAs each contribute roughly another 10%.

43. a. Since  $2^6 = 64 < 70 < 128 = 2^7$ , 7 classes are recommended. The interval should be at least  $(1,002.2 - 3.3)/7 = 142.7$ . Use 150 as a convenient value.

b. Based on the histogram, the majority of people have less than \$500,000 in their investment portfolio and may not have enough money for retirement. Merrill Lynch financial advisors need to promote the importance of investing for retirement in this age group.



45. a. Pie chart

b. 700, found by  $0.7(1,000)$

c. Yes,  $0.70 + 0.20 = 0.90$

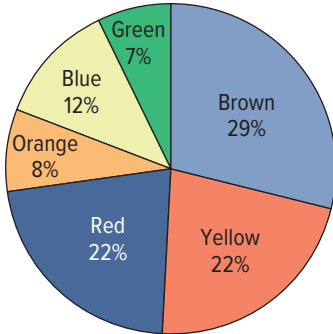
47. a.



- b. 24.0%, found by  $(40.0 + 23.9)/266$   
 c. 45.7%, found by  $(40.0 + 23.9)/(48.1 + 40.0 + 23.9 + 15.5 + 12.3)$

49.

M & M s



Brown, yellow, and red make up almost 75% of the candies. The other 25% is composed of blue, orange, and green.

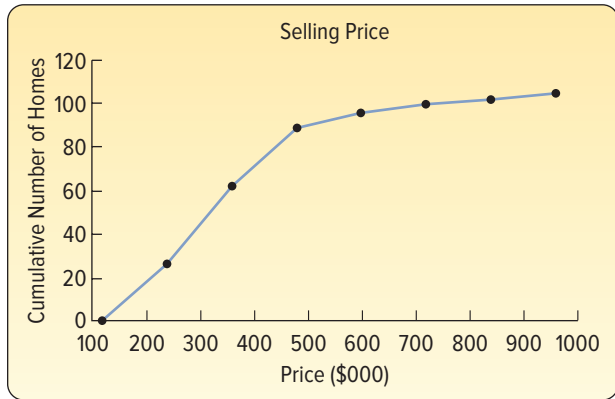
51. There are many choices and possibilities here. For example you could choose to start the first class at 160,000 rather than 120,000. The choice is yours!

$i \geq (919,480 - 167,962)/7 = 107,360$ . Use intervals of 120,000.

| Selling Price (000) | Frequency | Cumulative Frequency |
|---------------------|-----------|----------------------|
| 120 up to 240       | 26        | 26                   |
| 240 up to 360       | 36        | 62                   |
| 360 up to 480       | 27        | 89                   |
| 480 up to 600       | 7         | 96                   |
| 600 up to 720       | 4         | 100                  |
| 720 up to 840       | 2         | 102                  |
| 840 up to 960       | 1         | 105                  |

- a. Most homes (60%) sell between \$240,000 and \$480,000.  
 b. The typical price in the first class is \$180,000 and in the last class is \$900,000.

c.

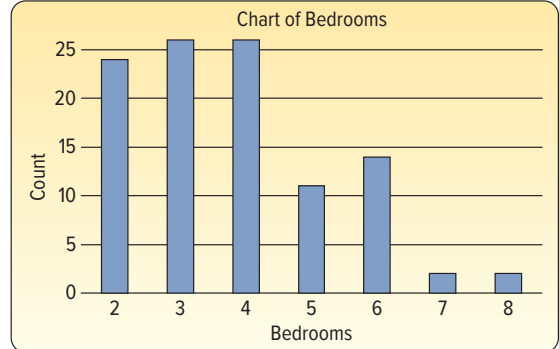


Fifty percent (about 52) of the homes sold for about \$320,000 or less.

The top 10% (about 90) of homes sold for at least \$520,000.

About 41% (about 41) of the homes sold for less than \$300,000.

d.

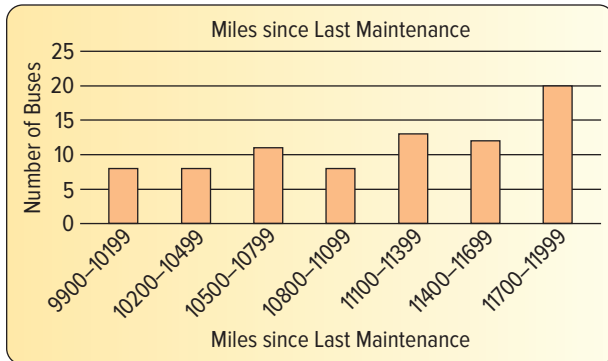


Two-, 3-, and 4-bedroom houses are most common with about 25 houses each. Seven- and 8-bedroom houses are rather rare.

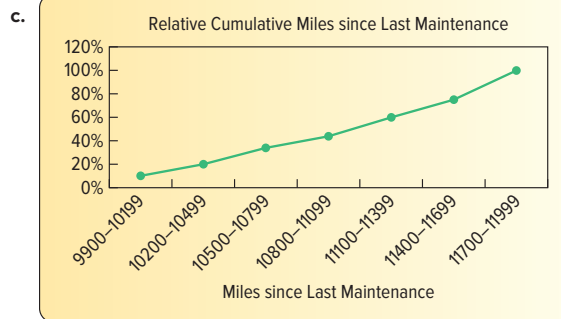
53. Since  $2^6 = 64 < 80 < 128 = 2^7$ , use 7 classes. The interval should be at least  $(11973 - 10000)/7 = 281$  miles. Use 300. The resulting frequency distribution is:

| Class             | f  |
|-------------------|----|
| 9900 up to 10299  | 8  |
| 10200 up to 10599 | 8  |
| 10500 up to 10899 | 11 |
| 10800 up to 11199 | 8  |
| 11110 up to 11499 | 13 |
| 11400 up to 11799 | 12 |
| 11700 up to 12099 | 20 |

- a. The typical amount driven, or the middle of the distribution, is about 11,100 miles. Based on the frequency distribution, the range is from 9,900 up to 12,000 miles.



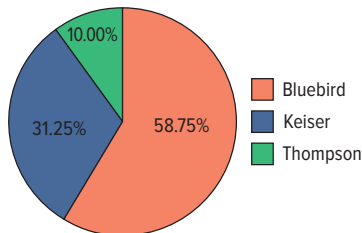
b. The distribution is somewhat “skewed” with a longer “tail” to the left and no outliers.



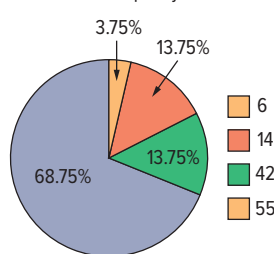
Forty percent of the buses were driven fewer than about 10,800 miles. About 30% of the 80 buses (about 24) were driven less than 10,500 miles.

d. The first diagram shows that Bluebird makes about 59% of the buses, Keiser about 31%, and Thompson only about 10%. The second chart shows that nearly 69% of the buses have 55 seats.

Pie Chart of Manufacturer



Bus Seat Capacity



### CHAPTER 3

- $\mu = 5.4$ , found by  $27/5$
- a.  $\bar{x} = 7.0$ , found by  $28/4$   
b.  $(5 - 7) + (9 - 7) + (4 - 7) + (10 - 7) = 0$
- $\bar{x} = 14.58$ , found by  $43.74/3$
- a. 15.4, found by  $154/10$   
b. Population parameter, since it includes all the salespeople at Midtown Ford

- a. \$54.55, found by  $\$1,091/20$   
b. A sample statistic—assuming that the power company serves more than 20 customers

11.  $\bar{x} = \frac{\sum x}{n}$  so

$$\sum x = \bar{x} \cdot n = (\$5,430)(30) = \$162,900$$

- a. No mode  
b. The given value would be the mode.  
c. 3 and 4 bimodal
- a. Mean = 3.583  
b. Median = 5  
c. Mode = 5
- a. Median = 2.9  
b. Mode = 2.9

19.  $\bar{x} = \frac{647}{11} = 58.82$

Median = 58, Mode = 58

Any of the three measures would be satisfactory.

21. a.  $\bar{x} = \frac{90.4}{12} = 7.53$

b. Median = 7.45. There are several modes: 6.5, 7.3, 7.8, and 8.7.

c.  $\bar{x} = \frac{33.8}{4} = 8.45$

Median = 8.7

About 1 percentage point higher in winter

23. \$41.73, found by  $\frac{300(\$41) + 400(\$39) + 400(\$45)}{300 + 400 + 400}$

25. \$17.75, found by  $(\$400 + \$750 + \$2,400)/200$

- a. 7, found by  $10 - 3$   
b. 6, found by  $30/5$   
c. 6.8, found by  $34/5$   
d. The difference between the highest number sold (10) and the smallest number sold (3) is 7. The typical squared deviation from 6 is 6.8.
- a. 30, found by  $54 - 24$   
b. 38, found by  $380/10$   
c. 74.4, found by  $744/10$   
d. The difference between 54 and 24 is 30. The average of the squared deviations from 38 is 74.4.

31.

| State      | Mean  | Median | Range |
|------------|-------|--------|-------|
| California | 33.10 | 34.0   | 32    |
| Iowa       | 24.50 | 25.0   | 19    |

The mean and median ratings were higher, but there was also more variation in California.

33. a. 5  
b.  $4.4$ , found by  $\frac{(8 - 5)^2 + (3 - 5)^2 + (7 - 5)^2 + (3 - 5)^2 + (4 - 5)^2}{5}$

35. a. \$2.77  
b. 1.26, found by  $\frac{(2.68 - 2.77)^2 + (1.03 - 2.77)^2 + (2.26 - 2.77)^2 + (4.30 - 2.77)^2 + (3.58 - 2.77)^2}{5}$

37. a. Range: 7.3, found by  $11.6 - 4.3$ . Arithmetic mean: 6.94, found by  $34.7/5$ . Variance: 6.5944, found by  $32.972/5$ . Standard deviation: 2.568, found by  $\sqrt{6.5944}$ .

b. Dennis has a higher mean return ( $11.76 > 6.94$ ). However, Dennis has greater spread in its returns on equity ( $16.89 > 6.59$ ).

39. a.  $\bar{x} = 4$   
 $s^2 = \frac{(7 - 4)^2 + \dots + (3 - 4)^2}{5 - 1} = \frac{22}{5 - 1} = 5.5$

b.  $s = 2.3452$

41. a.  $\bar{x} = 38$   

$$s^2 = \frac{(28 - 38)^2 + \dots + (42 - 38)^2}{10 - 1}$$

$$= \frac{744}{10 - 1} = 82.667$$
b.  $s = 9.0921$
43. a.  $\bar{x} = \frac{951}{10} = 95.1$   

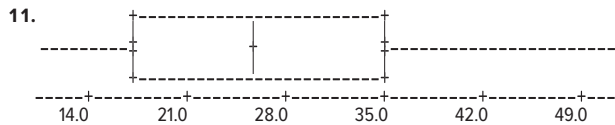
$$s^2 = \frac{(101 - 95.1)^2 + \dots + (88 - 95.1)^2}{10 - 1}$$

$$= \frac{1,112.9}{9} = 123.66$$
b.  $s = \sqrt{123.66} = 11.12$
45. About 69%, found by  $1 - 1/(1.8)^2$
47. a. About 95%  
b. 47.5%, 2.5%
49. a. Mean = 5, found by  $(6 + 4 + 3 + 7 + 5)/5$ .  
Median is 5, found by rearranging the values and selecting the middle value.  
b. Population, because all partners were included  
c.  $\Sigma(x - \mu) = (6 - 5) + (4 - 5) + (3 - 5) + (7 - 5) + (5 - 5) = 0$
51.  $\bar{x} = \frac{545}{16} = 34.06$   
Median = 37.50
53. The mean is 35.675, found by  $1,427/40$ . The median is 36, found by sorting the data and averaging the 20th and 21st observations.
55.  $\bar{x}_w = \frac{\$5.00(270) + \$6.50(300) + \$8.00(100)}{270 + 300 + 100} = \$6.12$
57.  $\bar{x}_w = \frac{15,300(4.5) + 10,400(3.0) + 150,600(10.2)}{176,300} = 9.28$
59. a. 55, found by  $72 - 17$   
b. 17.6245, found by the square root of  $2795.6/9$
61. a. This is a population because it includes all the public universities in Ohio.  
b. The mean is 25,165.4.  
c. The median is 20,595.  
d. The range is 60,560.  
e. The standard deviation is 16,344.9.
63. a. The mean is \$717.20, found by  $\$17,930/25$ . The median is \$717.00 and there are two modes, \$710 and \$722.  
b. The range is \$90, found by  $\$771 - \$681$ , and the standard deviation is \$24.87, found by the square root of  $14,850/24$ .  
c. From \$667.46 up to \$766.94, found by  $\$717.20 \pm 2(\$24.87)$
65.  $\bar{x} = \frac{273}{30} = 9.1$ , Median = 9
67. a. The mean team salary is \$121.12 million and the median is \$112.91 million. Since the distribution is skewed, the median value of \$112.91 million is more typical.  
b. The range is \$154.29 million, found by  $\$223.35$  million -  $\$69.06$  million. The population standard deviation is \$39.66. At least 95% of the team salaries are between \$74.97 million and \$233.61 million, found by  $\$154.29$  million plus or minus  $2(\$39.66$  million).

## CHAPTER 4

1. In a histogram, observations are grouped so their individual identity is lost. With a dot plot, the identity of each observation is maintained.
3. a. Dot plot  
b. 15  
c. 1, 7  
d. 2 and 3
5. Median = 53, found by  $(11 + 1)(\frac{1}{2})$ . ∴ 6th value in from lowest  
 $Q_1 = 49$ , found by  $(11 + 1)(\frac{1}{4})$ . ∴ 3rd value in from lowest  
 $Q_3 = 55$ , found by  $(11 + 1)(\frac{3}{4})$ . ∴ 9th value in from lowest

7. a.  $Q_1 = 33.25, Q_3 = 50.25$   
b.  $D_2 = 27.8, D_8 = 52.6$   
c.  $P_{67} = 47$
9. a. 800  
b.  $Q_1 = 500, Q_3 = 1,200$   
c. 700, found by  $1,200 - 500$   
d. Less than 200 or more than 1,800  
e. There are no outliers.  
f. The distribution is positively skewed.



The distribution is somewhat positively skewed. Note that the dashed line above 35 is longer than below 18.

13. a. The mean is 30.8, found by  $154/5$ . The median is 31.0, and the standard deviation is 3.96, found by

$$s = \sqrt{\frac{62.8}{4}} = 3.96$$

- b. -0.15, found by  $\frac{3(30.8 - 31.0)}{3.96}$

c.

| Salary | $\left(\frac{x - \bar{x}}{s}\right)$ | $\left(\frac{x - \bar{x}}{s}\right)^3$ |
|--------|--------------------------------------|--|
| 36     | 1.313131                             | 2.264250504                            |
| 26     | -1.212121                            | -1.780894343                           |
| 33     | 0.555556                             | 0.171467764                            |
| 28     | -0.707071                            | -0.353499282                           |
| 31     | 0.050505                             | 0.000128826                            |
|        |                                      | 0.301453469                            |

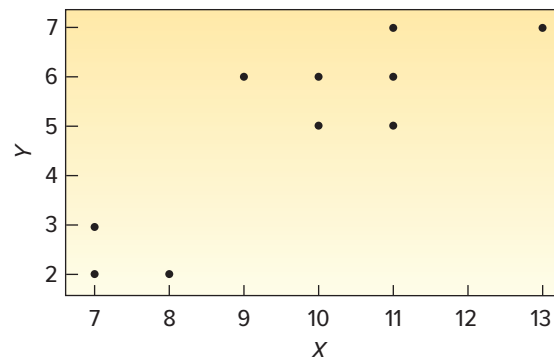
0.125, found by  $[5/(4 \times 3)] \times 0.301$

15. a. The mean is 21.93, found by  $328.9/15$ . The median is 15.8, and the standard deviation is 21.18, found by

$$s = \sqrt{\frac{6,283}{14}} = 21.18$$

- b. 0.868, found by  $[3(21.93 - 15.8)]/21.18$   
c. 2.444, found by  $[15/(14 \times 13)] \times 29.658$

17. Scatter Diagram of Y versus X

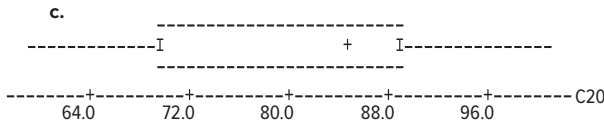


There is a positive relationship between the variables.

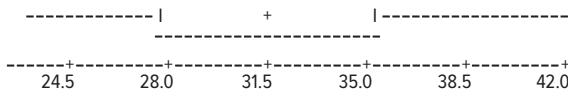
19. a. Both variables are nominal scale.  
b. Contingency table  
c. Men are about twice as likely to order a dessert. From the table, 32% of the men ordered dessert, but only 15% of the women.

21. a. Dot plot                      b. 15                      c. 5

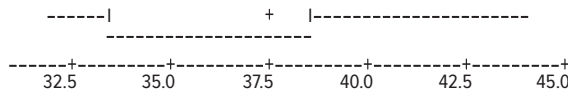
23. a.  $L_{50} = (20 + 1)\frac{50}{100} = 10.50$   
 $\text{Median} = \frac{83.7 + 85.6}{2} = 84.65$   
 $L_{25} = (21)(.25) = 5.25$   
 $Q_1 = 66.6 + .25(72.9 - 66.6) = 68.175$   
 $L_{75} = 21(.75) = 15.75$   
 $Q_3 = 87.1 + .75(90.2 - 87.1) = 89.425$   
 b.  $L_{26} = 21(.26) = 5.46$   
 $P_{26} = 66.6 + .46(72.9 - 66.6) = 69.498$   
 $L_{83} = 21(.83) = 17.43$   
 $P_{83} = 93.3 + .43(98.6 - 93.3) = 95.579$



25. a.  $Q_1 = 26.25$ ,  $Q_3 = 35.75$ , Median = 31.50



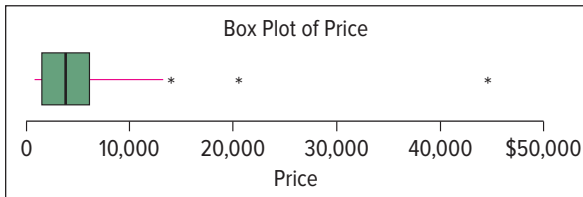
- b.  $Q_1 = 33.25$ ,  $Q_3 = 38.75$ , Median = 37.50



- c. The median time for public transportation is about 6 minutes less. There is more variation in public transportation. The difference between  $Q_1$  and  $Q_3$  is 9.5 minutes for public transportation and 5.5 minutes for private transportation.

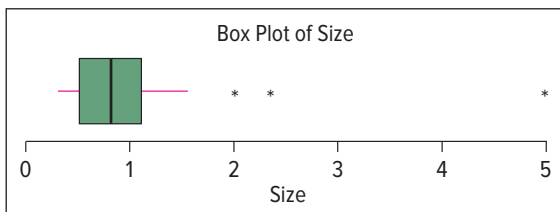
27. The distribution is positively skewed. The first quartile is about \$20 and the third quartile is about \$90. There is one outlier located at \$255. The median is about \$50.

29. a.



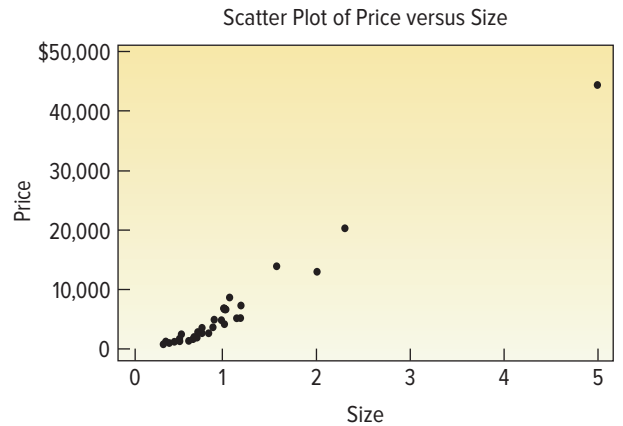
Median is 3,733. First quartile is 1,478. Third quartile is 6,141. So prices over 13,135.5, found by  $6,141 + 1.5 \times (6,141 - 1,478)$ , are outliers. There are three (13,925; 20,413; and 44,312).

- b.



Median is 0.84. First quartile is 0.515. Third quartile is 1.12. So sizes over 2.0275, found by  $1.12 + 1.5(1.12 - 0.515)$ , are outliers. There are three (2.03; 2.35; and 5.03).

- c.



There is a direct association between them. The first observation is larger on both scales.

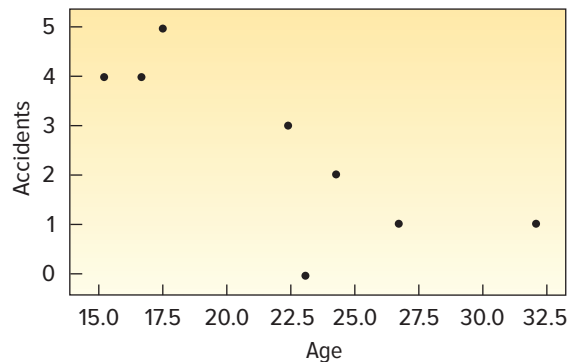
- d.

| Shape/<br>Cut |         |      |       |         |       | Ultra | All |
|---------------|---------|------|-------|---------|-------|-------|-----|
|               | Average | Good | Ideal | Premium | Ideal |       |     |
| Emerald       | 0       | 0    | 1     | 0       | 0     | 1     |     |
| Marquise      | 0       | 2    | 0     | 1       | 0     | 3     |     |
| Oval          | 0       | 0    | 0     | 1       | 0     | 1     |     |
| Princess      | 1       | 0    | 2     | 2       | 0     | 5     |     |
| Round         | 1       | 3    | 3     | 13      | 3     | 23    |     |
| Total         | 2       | 5    | 6     | 17      | 3     | 33    |     |

The majority of the diamonds are round (23). Premium cut is most common (17). The Round Premium combination occurs most often (13).

31.  $sk = 0.065$  or  $sk = \frac{3(7.7143 - 8.0)}{3.9036} = -0.22$

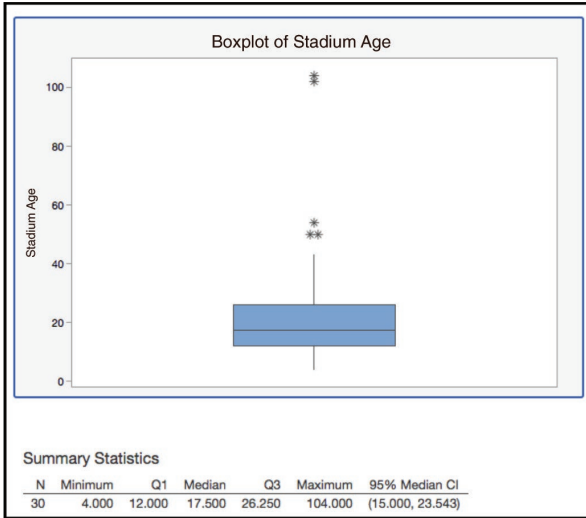
33. Scatter Plot of Accidents versus Age



As age increases, the number of accidents decreases.

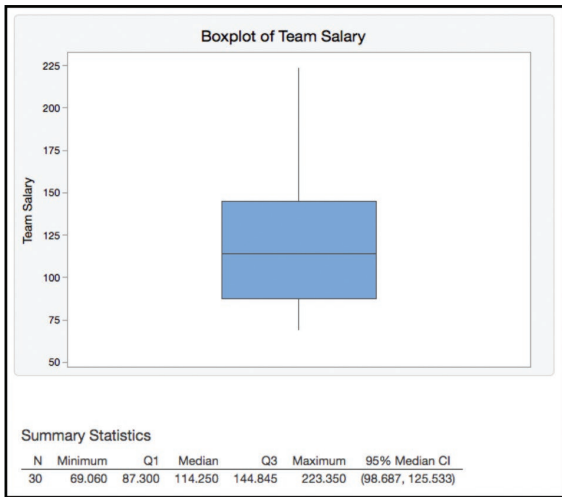
35. a. 139,340,000  
 b. 5.4% unemployed, found by  $(7,523/139,340)100$   
 c. Men = 5.64%  
 Women = 5.12%

37. a. Box plot of age assuming the current year is 2016.



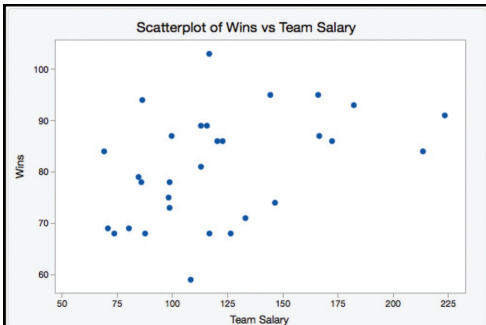
Distribution of stadium age is highly skewed to the right. Any stadium older than  $47.625$  years ( $Q3 + 1.5(Q3 - Q1) = 26.25 + 1.5(26.25 - 12)$ ) is an outlier. Boston, Chicago Cubs, LA Dodgers, Oakland, and LA Angels.

b.



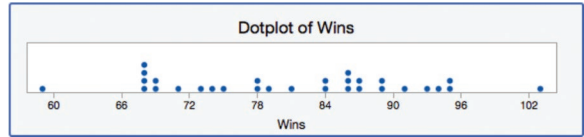
The first quartile is \$87.3 million and the third is \$144.35 million. Outliers are greater than  $(Q3 + 1.5(Q3 - Q1))$  or  $144.35 + 1.5(144.35 - 87.3) = \$229.925$  million. The distribution is positively skewed. However in 2016, there were no outliers.

c.



Higher salaries do not necessarily lead to more wins.

d.



The distribution is fairly uniform between 59 and 103.

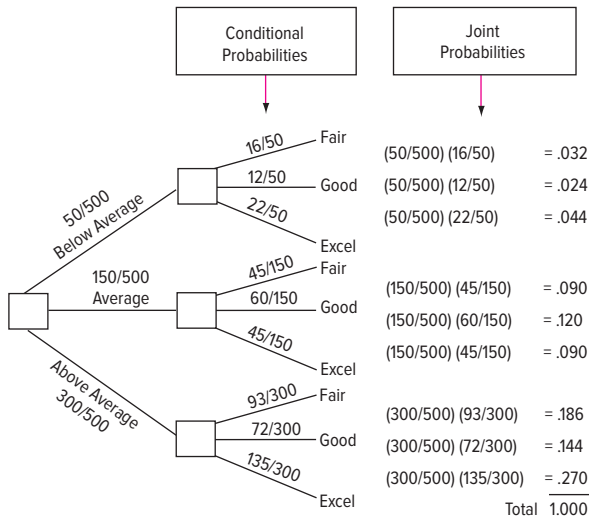
## CHAPTER 5

1.

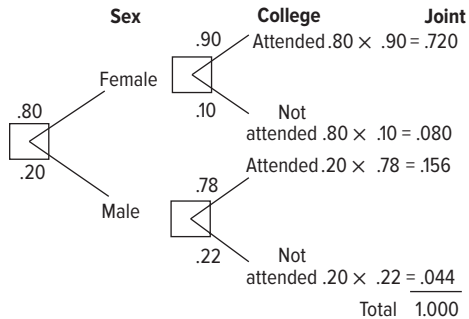
| Outcome | Person |   |
|---------|--------|---|
|         | 1      | 2 |
| 1       | A      | A |
| 2       | A      | F |
| 3       | F      | A |
| 4       | F      | F |

3. a.  $.176$ , found by  $\frac{6}{34}$       b. Empirical
5. a. Empirical  
b. Classical  
c. Classical  
d. Empirical, based on seismological data
7. a. The survey of 40 people about environmental issues  
b. 26 or more respond yes, for example.  
c.  $10/40 = .25$   
d. Empirical  
e. The events are not equally likely, but they are mutually exclusive.
9. a. Answers will vary. Here are some possibilities: 1234, 5694, 6722, 9999.  
b.  $(1/10)^4$   
c. Classical
11.  $P(A \text{ or } B) = P(A) + P(B) = .30 + .20 = .50$   
 $P(\text{neither}) = 1 - .50 = .50$
13. a.  $102/200 = .51$   
b.  $.49$ , found by  $61/200 + 37/200 = .305 + .185$ . Special rule of addition.
15.  $P(\text{above } C) = .25 + .50 = .75$
17.  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = .20 + .30 - .15 = .35$
19. When two events are mutually exclusive, it means that if one occurs, the other event cannot occur. Therefore, the probability of their joint occurrence is zero.
21. Let  $A$  denote the event the fish is green and  $B$  be the event the fish is male.  
a.  $P(A) = 80/140 = 0.5714$   
b.  $P(B) = 60/140 = 0.4286$   
c.  $P(A \text{ and } B) = 36/140 = 0.2571$   
d.  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = 80/140 + 60/140 - 36/140 = 104/140 = 0.7429$
23.  $P(A \text{ and } B) = P(A) \times P(B|A) = .40 \times .30 = .12$
25.  $.90$ , found by  $(.80 + .60) - .5$   
 $.10$ , found by  $(1 - .90)$
27. a.  $P(A_1) = 3/10 = .30$   
b.  $P(B_1|A_2) = 1/3 = .33$   
c.  $P(B_2 \text{ and } A_3) = 1/10 = .10$
29. a. A contingency table  
b.  $.27$ , found by  $300/500 \times 135/300$

c. The tree diagram would appear as:



31. a. Out of all 545 students, 171 prefer skiing. So the probability is  $171/545$ , or 0.3138.  
 b. Out of all 545 students, 155 are in junior college. Thus, the probability is  $155/545$ , or 0.2844.  
 c. Out of 210 four-year students, 70 prefer ice skating. So the probability is  $70/210$ , or 0.3333.  
 d. Out of 211 students who prefer snowboarding, 68 are in junior college. So the probability is  $68/211$ , or 0.3223.  
 e. Out of 180 graduate students, 74 prefer skiing and 47 prefer ice skating. So the probability is  $(74 + 47)/180 = 121/180$ , or 0.6722.
33. a. 78,960,960  
 b. 840, found by  $(7)(6)(5)(4)$ . That is  $7!/3!$   
 c. 10, found by  $5!/3!2!$
35. 210, found by  $(10)(9)(8)(7)(4)(3)(2)$
37. 120, found by  $5!$
39.  $(4)(8)(3) = 96$  combinations
41. a. Asking teenagers to compare their reactions to a newly developed soft drink.  
 b. Answers will vary. One possibility is more than half of the respondents like it.
43. Subjective
45. a.  $4/9$ , found by  $(2/3) \cdot (2/3)$   
 b.  $3/4$ , because  $(3/4) \cdot (2/3) = 0.5$
47. a. .8145, found by  $(.95)^4$   
 b. Special rule of multiplication  
 c.  $P(A \text{ and } B \text{ and } C \text{ and } D) = P(A) \times P(B) \times P(C) \times P(D)$
49. a. .08, found by  $.80 \times .10$   
 b. No; 90% of females attended college, 78% of males  
 c.



- d. Yes, because all the possible outcomes are shown on the tree diagram.

51. a. 0.57, found by  $57/100$   
 b. 0.97, found by  $(57/100) + (40/100)$   
 c. Yes, because an employee cannot be both.  
 d. 0.03, found by  $1 - 0.97$
53. a.  $1/2$ , found by  $(2/3)(3/4)$   
 b.  $1/12$ , found by  $(1/3)(1/4)$   
 c.  $11/12$ , found by  $1 - 1/12$
55. a. 0.9039, found by  $(0.98)^5$   
 b. 0.0961, found by  $1 - 0.9039$
57. a. 0.0333, found by  $(4/10)(3/9)(2/8)$   
 b. 0.1667, found by  $(6/10)(5/9)(4/8)$   
 c. 0.8333, found by  $1 - 0.1667$   
 d. Dependent
59. a. 0.3818, found by  $(9/12)(8/11)(7/10)$   
 b. 0.6182, found by  $1 - 0.3818$
61. a.  $P(S) \cdot P(R|S) = .60(.85) = 0.51$   
 b.  $P(S) \cdot P(PR|S) = .60(1 - .85) = 0.09$
63. a.  $P(\text{not perfect}) = P(\text{bad sector}) + P(\text{defective})$   

$$= \frac{112}{1,000} + \frac{31}{1,000} = .143$$
  
 b.  $P(\text{defective} | \text{not perfect}) = \frac{.031}{.143} = .217$
65. a.  $0.1 + 0.02 = 0.12$   
 b.  $1 - 0.12 = 0.88$   
 c.  $(0.88)^3 = 0.6815$   
 d.  $1 - .6815 = 0.3185$
67. Yes, 256 is found by  $2^8$ .
69. .9744, found by  $1 - (.40)^4$
71. a. 0.193, found by  $.15 + .05 - .0075 = .193$   
 b. .0075, found by  $(.15)(.05)$
73. a.  $P(F \text{ and } >60) = .25$ , found by solving with the general rule of multiplication:  $P(F) \cdot P(>60|F) = (.5)(.5)$   
 b. 0  
 c. .3333, found by  $1/3$
75.  $26^4 = 456,976$
77. There are  $10!$  or 3,628,800 possible matches. The probability of randomly matching all 10 correctly is  $(1/3,628,800)$  or 0.00000028.
79. 0.512, found by  $(0.8)^3$
81. .525, found by  $1 - (.78)^3$
83. a.

| Winning Season | Low Attendance | Moderate Attendance | High Attendance | Total |
|----------------|----------------|---------------------|-----------------|-------|
| No             | 7              | 6                   | 1               | 14    |
| Yes            | 1              | 9                   | 6               | 16    |
| Total          | 8              | 15                  | 7               | 30    |

1. 0.5333, found by  $16/30$   
 2. 0.5667, found by  $16/30 + 5/30 - 4/30 = 17/30$   
 3. 0.8571, found by  $6/7$   
 4. 0.0333, found by  $1/30$

b.

|       | Losing Season | Winning Season | Total |
|-------|---------------|----------------|-------|
| New   | 8             | 8              | 16    |
| Old   | 6             | 8              | 14    |
| Total | 14            | 16             | 30    |

1. 0.4667, found by  $14/30$   
 2. 0.2667, found by  $8/30$   
 3. 0.8000, found by  $16/30 + 16/30 - 8/30 = 24/30$

## CHAPTER 6

1. Mean = 1.3, variance = .81, found by:  

$$\mu = 0(.20) + 1(.40) + 2(.30) + 3(.10) = 1.3$$

$$\sigma^2 = (0 - 1.3)^2(.2) + (1 - 1.3)^2(.4) + (2 - 1.3)^2(.3) + (3 - 1.3)^2(.1)$$

$$= .81$$



3. Mean = 14.5, variance = 27.25, found by:  
 $\mu = 5(.1) + 10(.3) + 15(.2) + 20(.4) = 14.5$   
 $\sigma^2 = (5 - 14.5)^2(.1) + (10 - 14.5)^2(.3)$   
 $+ (15 - 14.5)^2(.2) + (20 - 14.5)^2(.4)$   
 $= 27.25$

5. a.

| Calls, $x$ | Frequency | $P(x)$ | $xP(x)$ | $(x - \mu)^2 P(x)$ |
|------------|-----------|--------|---------|--------------------|
| 0          | 8         | .16    | 0       | .4624              |
| 1          | 10        | .20    | .20     | .0980              |
| 2          | 22        | .44    | .88     | .0396              |
| 3          | 9         | .18    | .54     | .3042              |
| 4          | 1         | .02    | .08     | .1058              |
|            | 50        |        | 1.70    | 1.0100             |

- b. Discrete distribution, because only certain outcomes are possible.  
 c.  $\mu = \Sigma x \cdot P(x) = 1.70$   
 d.  $\sigma = \sqrt{1.01} = 1.005$

7.

| Amount | $P(x)$ | $xP(x)$ | $(x - \mu)^2 P(x)$ |
|--------|--------|---------|--------------------|
| 10     | .50    | 5       | 60.50              |
| 25     | .40    | 10      | 6.40               |
| 50     | .08    | 4       | 67.28              |
| 100    | .02    | 2       | 124.82             |
|        |        | 21      | 259.00             |

- a.  $\mu = \Sigma xP(x) = 21$   
 b.  $\sigma^2 = \Sigma (x - \mu)^2 P(x) = 259$   
 $\sigma = \sqrt{259} = 16.093$

9. a.  $P(2) = \frac{4!}{2!(4-2)!} (.25)^2 (.75)^{4-2} = .2109$

b.  $P(3) = \frac{4!}{3!(4-3)!} (.25)^3 (.75)^{4-3} = .0469$

11. a.

| $x$ | $P(x)$ |
|-----|--------|
| 0   | .064   |
| 1   | .288   |
| 2   | .432   |
| 3   | .216   |

- b.  $\mu = 1.8$   
 $\sigma^2 = 0.72$   
 $\sigma = \sqrt{0.72} = .8485$

13. a. .2668, found by  $P(2) = \frac{9!}{(9-2)!2!} (.3)^2 (.7)^7$

b. .1715, found by  $P(4) = \frac{9!}{(9-4)!4!} (.3)^4 (.7)^5$

c. .0404, found by  $P(0) = \frac{9!}{(9-0)!0!} (.3)^0 (.7)^9$

15. a. .2824, found by  $P(0) = \frac{12!}{(12-0)!0!} (.1)^0 (.9)^{12}$

b. .3766, found by  $P(1) = \frac{12!}{(12-1)!1!} (.1)^1 (.9)^{11}$

c. .2301, found by  $P(2) = \frac{12!}{(12-2)!2!} (.1)^2 (.9)^{10}$

- d.  $\mu = 1.2$ , found by  $12(.1)$   
 $\sigma = 1.0392$ , found by  $\sqrt{1.08}$

17. a. 0.1858, found by  $\frac{15!}{2!13!} (0.23)^2 (0.77)^{13}$

b. 0.1416, found by  $\frac{15!}{5!10!} (0.23)^5 (0.77)^{10}$

c. 3.45, found by  $(0.23)(15)$

19. a. 0.296, found by using Appendix B.1 with  $n$  of 8,  $\pi$  of 0.30, and  $x$  of 2

b.  $P(x \leq 2) = 0.058 + 0.198 + 0.296 = 0.552$

c. 0.448, found by  $P(x \geq 3) = 1 - P(x \leq 2) = 1 - 0.552$

21. a. 0.387, found from Appendix B.1 with  $n$  of 9,  $\pi$  of 0.90, and  $x$  of 9

b.  $P(x < 5) = 0.001$

c. 0.992, found by  $1 - 0.008$

d. 0.947, found by  $1 - 0.053$

23. a.  $\mu = 10.5$ , found by  $15(0.7)$  and  $\sigma = \sqrt{15(0.7)(0.3)} = 1.7748$

b. 0.2061, found by  $\frac{15!}{10!5!} (0.7)^{10} (0.3)^5$

c. 0.4247, found by  $0.2061 + 0.2186$

d. 0.5154, found by

$0.2186 + 0.1700 + 0.0916 + 0.0305 + 0.0047$

25. a. .6703      b. .3297

27. a. .0613      b. .0803

29.  $\mu = 6$

$P(x \geq 5) = 1 - (.0025 + .0149 + .0446 + .0892 + .1339)$   
 $= .7149$

31. A random variable is an outcome that results from a chance experiment. A probability distribution also includes the likelihood of each possible outcome.

33.  $\mu = \$1,000(.25) + \$2,000(.60) + \$5,000(.15) = \$2,200$

$\sigma^2 = (1,000 - 2,200)^2(.25) + (2,000 - 2,200)^2(.60) + (5,000 - 2,200)^2(.15)$   
 $= 1,560,000$

35.  $\mu = 12(.25) + \dots + 15(.1) = 13.2$

$\sigma^2 = (12 - 13.2)^2(.25) + \dots + (15 - 13.2)^2(.10) = 0.86$

$\sigma = \sqrt{0.86} = .927$

37. a.  $10(.35) = 3.5$

b.  $P(x = 4) = {}_{10}C_4 (.35)^4 (.65)^6 = 210(.0150)(.0754) = .2375$

c.  $P(x \geq 4) = {}_{10}C_4 (.35)^4 (.65)^{10-4} + \dots + .0000 = .4862$

39. a. 6, found by  $0.4 \times 15$

b. 0.0245, found by  $\frac{15!}{10!5!} (0.4)^{10} (0.6)^5$

c. 0.0338, found by

$0.0245 + 0.0074 + 0.0016 + 0.0003 + 0.0000$

d. 0.0093, found by  $0.0338 - 0.0245$

41. a.  $\mu = 20(0.075) = 1.5$

$\sigma = \sqrt{20(0.075)(0.925)} = 1.1779$

b. 0.2103, found by  $\frac{20!}{0!20!} (0.075)^0 (0.925)^{20}$

c. 0.7897, found by  $1 - 0.2103$

43. a. 0.1311, found by  $\frac{16!}{4!12!} (0.15)^4 (0.85)^{12}$

b. 2.4, found by  $(0.15)(16)$

c. 0.2100, found by  $1 - 0.0743 - 0.2097 - 0.2775 - 0.2285$

45. 0.2784, found by  $0.1472 + 0.0811 + 0.0348 + 0.0116 + 0.0030 + 0.0006 + 0.0001 + 0.0000$

47. a.

|   |        |    |        |
|---|--------|----|--------|
| 0 | 0.0002 | 7  | 0.2075 |
| 1 | 0.0019 | 8  | 0.1405 |
| 2 | 0.0116 | 9  | 0.0676 |
| 3 | 0.0418 | 10 | 0.0220 |
| 4 | 0.1020 | 11 | 0.0043 |
| 5 | 0.1768 | 12 | 0.0004 |
| 6 | 0.2234 |    |        |

b.  $\mu = 12(0.52) = 6.24$        $\sigma = \sqrt{12(0.52)(0.48)} = 1.7307$

c. 0.1768

d. 0.3343, found by

$0.0002 + 0.0019 + 0.0116 + 0.0418 + 0.1020 + 0.1768$

49. a. .0183      b. .1954

c. .6289      d. .5665

51. a. 0.1733, found by  $\frac{(3.1)^4 e^{-3.1}}{4!}$

b. 0.0450, found by  $\frac{(3.1)^0 e^{-3.1}}{0!}$

c. 0.9550, found by  $1 - 0.0450$

$$53. \mu = n\pi = 23 \left( \frac{2}{113} \right) = .407$$

$$P(2) = \frac{(.407)^2 e^{-.407}}{2!} = 0.0551$$

$$P(0) = \frac{(.407)^0 e^{-.407}}{0!} = 0.6656$$

$$55. \text{ Let } \mu = n\pi = 155(1/3,709) = 0.042$$

$$P(4) = \frac{0.042^4 e^{-0.042}}{4!} = 0.00000012$$

Very unlikely!

$$57. \text{ a. } \mu = n\pi = 15(.67) = 10.05$$

$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{15(.67)(.33)} = 1.8211$$

$$\text{ b. } P(8) = {}_{15}C_8 (.67)^8 (.33)^7 = 6435(.0406)(.000426) = .1114$$

$$\text{ c. } P(x \geq 8) = .1114 + .1759 + \dots + .0025 = .9163$$

59. The mean number of home runs per game is 2.3. The average season home runs per team is 187. Then  $(187 \times 2)/162 = 2.3$ .

$$\text{ a. } P(0) = \frac{2.3^0 e^{-2.3}}{0!} = 0.1003$$

$$\text{ b. } P(2) = \frac{2.3^2 e^{-2.3}}{2!} = 0.2652$$

$$\text{ c. } P(x \geq 4) = 0.1981, \text{ found by } 1 - (0.1169 + 0.0538 + 0.0206 + 0.0068)$$

## CHAPTER 7

$$1. \text{ a. } b = 10, \sigma = 6 \qquad \text{ b. } \mu = \frac{6+10}{2} = 8$$

$$\text{ c. } \sigma = \sqrt{\frac{(10-6)^2}{12}} = 1.1547$$

$$\text{ d. Area} = \frac{1}{(10-6)} \cdot \frac{(10-6)}{1} = 1$$

$$\text{ e. } P(x > 7) = \frac{1}{(10-6)} \cdot \frac{10-7}{1} = \frac{3}{4} = .75$$

$$\text{ f. } P(7 \leq x \leq 9) = \frac{1}{(10-6)} \cdot \frac{(9-7)}{1} = \frac{2}{4} = .50$$

$$3. \text{ a. } 0.30, \text{ found by } (30-27)/(30-20)$$

$$\text{ b. } 0.40, \text{ found by } (24-20)/(30-20)$$

$$5. \text{ a. } \sigma = 0.5, b = 3.00$$

$$\text{ b. } \mu = \frac{0.5 + 3.00}{2} = 1.75$$

$$\sigma = \sqrt{\frac{(3.00 - .50)^2}{12}} = .72$$

$$\text{ c. } P(x < 1) = \frac{1}{(3.0 - 0.5)} \cdot \frac{1 - .5}{1} = \frac{.5}{2.5} = 0.2$$

$$\text{ d. } 0, \text{ found by } \frac{1}{(3.0 - 0.5)} \cdot \frac{(1.0 - 1.0)}{1}$$

$$\text{ e. } P(x > 1.5) = \frac{1}{(3.0 - 0.5)} \cdot \frac{3.0 - 1.5}{1} = \frac{1.5}{2.5} = 0.6$$

7. The actual shape of a normal distribution depends on its mean and standard deviation. Thus, there is a normal distribution, and an accompanying normal curve, for a mean of 7 and a standard deviation of 2. There is another normal curve for a mean of \$25,000 and a standard deviation of \$1,742, and so on.

$$9. \text{ a. } 490 \text{ and } 510, \text{ found by } 500 \pm 1(10)$$

$$\text{ b. } 480 \text{ and } 520, \text{ found by } 500 \pm 2(10)$$

$$\text{ c. } 470 \text{ and } 530, \text{ found by } 500 \pm 3(10)$$

$$11. z_{\text{Rob}} = \frac{\$50,000 - \$60,000}{\$5,000} = -2$$

$$z_{\text{Rachel}} = \frac{\$50,000 - \$35,000}{\$8,000} = 1.875$$

Adjusting for their industries, Rob is well below average and Rachel well above.

$$13. \text{ a. } 1.25, \text{ found by } z = \frac{25 - 20}{4.0} = 1.25$$

$$\text{ b. } 0.3944, \text{ found in Appendix B.3}$$

$$\text{ c. } 0.3085, \text{ found by } z = \frac{18 - 20}{2.5} = -0.5$$

Find 0.1915 in Appendix B.3 for  $z = -0.5$ , then  $0.5000 - 0.1915 = 0.3085$

$$15. \text{ a. } 0.3413, \text{ found by } z = \frac{\$24 - \$20.50}{\$3.50} = 1.00, \text{ then find } 0.3413$$

in Appendix B.3 for  $z = 1$

$$\text{ b. } 0.1587, \text{ found by } 0.5000 - 0.3413 = 0.1587$$

$$\text{ c. } 0.3336, \text{ found by } z = \frac{\$19.00 - \$20.50}{\$3.50} = -0.43$$

Find 0.1664 in Appendix B.3, for  $z = -0.43$ , then  $0.5000 - 0.1664 = 0.3336$

17. a. 0.8276: First find  $z = -1.5$ , found by  $(44 - 50)/4$  and  $z = 1.25 = (55 - 50)/4$ . The area between  $-1.5$  and 0 is 0.4332 and the area between 0 and 1.25 is 0.3944, both from Appendix B.3. Adding the two areas, we find that  $0.4332 + 0.3944 = 0.8276$ .

$$\text{ b. } 0.1056, \text{ found by } 0.5000 - .3944, \text{ where } z = 1.25$$

c. 0.2029: Recall that the area for  $z = 1.25$  is 0.3944, and the area for  $z = 0.5$ , found by  $(52 - 50)/4$ , is 0.1915. Then subtract  $0.3944 - 0.1915$  and find 0.2029.

19. a. 0.2514: Begin by using formula (7-5) to find the  $z$  value for \$3,100, which is  $(3,100 - 2,800)/450$ , or 0.67. Then see Appendix B.3 to find the area between 0 and 0.67, which is 0.2486. Finally, since the area of interest is beyond 0.67, subtract that probability from 0.5000. The result is  $0.5000 - 0.2486$ , or 0.2514.

b. 0.1908: Use formula (7-5) to find the  $z$  value for \$3,500, which is  $(3,500 - 2,800)/450$ , or 1.56. Then see Appendix B.3 for the area under the standard normal curve. That probability is 0.4406. Since the two points (1.56 and 0.66) are on the same side of the mean, subtract the smaller probability from the larger. The result is  $0.4406 - 0.2486 = 0.1920$ .

c. 0.8294: Use formula (7-5) to find the  $z$  value for \$2,250, which is  $-1.22$ , found by  $(2,250 - 2,800)/450$ . The corresponding area is 0.3888. Since 1.56 and  $-1.22$  are on different sides of the mean, add the corresponding probabilities. Thus, we find  $0.3888 + 0.4406 = 0.8294$ .

$$21. \text{ a. } 0.0764, \text{ found by } z = (20 - 15)/3.5 = 1.43, \text{ then } 0.5000 - 0.4236 = 0.0764$$

$$\text{ b. } 0.9236, \text{ found by } 0.5000 + 0.4236, \text{ where } z = 1.43$$

$$\text{ c. } 0.1185, \text{ found by } z = (12 - 15)/3.5 = -0.86.$$

The area under the curve is 0.3051, then  $z = (10 - 15)/3.5 = -1.43$ . The area is 0.4236. Finally,  $0.4236 - 0.3051 = 0.1185$ .

23.  $x = 56.60$ , found by adding 0.5000 (the area left of the mean) and then finding a  $z$  value that forces 45% of the data to fall inside the curve. Solving for  $x$ :  $1.65 = (x - 50)/4$ , so  $x = 56.60$ .

$$25. \$1,630, \text{ found by } \$2,100 - 1.88(\$250)$$

27. a. 214.8 hours: Find a  $z$  value where 0.4900 of area is between 0 and  $z$ . That value is  $z = 2.33$ . Then solve for  $x$ :  $2.33 = (x - 195)/8.5$ , so  $x = 214.8$  hours.

b. 270.2 hours: Find a  $z$  value where 0.4900 of area is between 0 and  $(-z)$ . That value is  $z = -2.33$ . Then solve for  $x$ :  $-2.33 = (x - 290)/8.5$ , so  $x = 270.2$  hours.

$$29. 41.7\%, \text{ found by } 12 + 1.65(18)$$

$$31. \text{ a. } \mu = \frac{11.96 + 12.05}{2} = 12.005$$

$$\text{ b. } \sigma = \sqrt{\frac{(12.05 - 11.96)^2}{12}} = .0260$$

$$\text{ c. } P(x < 12) = \frac{1}{(12.05 - 11.96)} \cdot \frac{12.00 - 11.96}{1} = \frac{.04}{.09} = .44$$

$$d. P(x > 11.98) = \frac{1}{(12.05 - 11.96)} \left( \frac{12.05 - 11.98}{1} \right) \\ = \frac{.07}{.09} = .78$$

e. All cans have more than 11.00 ounces, so the probability is 100%.

$$33. a. \mu = \frac{4 + 10}{2} = 7$$

$$b. \sigma = \sqrt{\frac{(10 - 4)^2}{12}} = 1.732$$

$$c. P(x < 6) = \frac{1}{(10 - 4)} \cdot \left( \frac{6 - 4}{1} \right) = \frac{2}{6} = .33$$

$$d. P(x > 5) = \frac{1}{(10 - 4)} \cdot \left( \frac{10 - 5}{1} \right) = \frac{5}{6} = .83$$

35. a.  $-0.4$  for net sales, found by  $(170 - 180)/25$ .  $2.92$  for employees, found by  $(1,850 - 1,500)/120$ .

b. Net sales is  $0.4$  standard deviation below the mean. Employees is  $2.92$  standard deviation above the mean.

c.  $65.54\%$  of the aluminum fabricators have greater net sales compared with Clarion, found by  $0.1554 + 0.5000$ . Only  $0.18\%$  have more employees than Clarion, found by  $0.5000 - 0.4982$ .

37. a.  $0.5000$ , because  $z = \frac{430 - 890}{90} = -5.11$

b.  $0.2514$ , found by  $0.5000 - 0.2486$

c.  $0.6374$ , found by  $0.2486 + 0.3888$

d.  $0.3450$ , found by  $0.3888 - 0.0438$

39. a.  $0.3015$ , found by  $0.5000 - 0.1985$

b.  $0.2579$ , found by  $0.4564 - 0.1985$

c.  $0.0011$ , found by  $0.5000 - 0.4989$

d.  $1,818$ , found by  $1,280 + 1.28(420)$

41. a.  $90.82\%$ : First find  $z = 1.33$ , found by  $(40 - 34)/4.5$ . The area between  $0$  and  $1.33$  is  $0.4082$  hours/week for women. Then add  $0.5000$  and  $0.4082$  and find  $0.9082$ , or  $90.82\%$ .

b.  $78.23\%$ : First find  $z = -0.78$ , found by  $(25 - 29)/5.1$ . The area between  $0$  and  $(-0.78)$  is  $0.2823$ . Then add  $0.5000$  and  $0.2823$  and find  $0.7823$ , or  $78.23\%$ .

c. Find a  $z$  value where  $0.4900$  of the area is between  $0$  and  $z$ . That value is  $2.33$ . Then solve for  $x$ :  $2.33 = (x - 34)/4.5$ , so  $x = 44.5$  hours/week for women.

$40.9$  hours/week for men:  $2.33 = (x - 29)/5.1$ , so  $x = 40.9$  hours/week.

43. About  $4,099$  units, found by solving for  $x$ .  $1.65 = (x - 4,000)/60$

45. a.  $15.39\%$ , found by  $(8 - 10.3)/2.25 = -1.02$ , then  $0.5000 - 0.3461 = 0.1539$ .

b.  $17.31\%$ , found by:

$$z = (12 - 10.3)/2.25 = 0.76. \text{ Area is } 0.2764.$$

$$z = (14 - 10.3)/2.25 = 1.64. \text{ Area is } 0.4495.$$

The area between  $12$  and  $14$  is  $0.1731$ , found by  $0.4495 - 0.2764$ .

c. On  $99.73\%$  of the days, returns are between  $3.55$  and  $17.05$ , found by  $10.3 \pm 3(2.25)$ . Thus, the chance of less than  $3.55$  returns is rather remote.

47. a.  $21.19\%$ , found by  $z = (9.00 - 9.20)/0.25 = -0.80$ , so  $0.5000 - 0.2881 = 0.2119$

b. Increase the mean.  $z = (9.00 - 9.25)/0.25 = -1.00$ ,  $P = 0.5000 - 0.3413 = 0.1587$

$$\text{Reduce the standard deviation. } \sigma = (9.00 - 9.20)/0.15 = -1.33; P = 0.5000 - 0.4082 = 0.0918$$

Reducing the standard deviation is better because a smaller percent of the hams will be below the limit.

49. a.  $z = (60 - 52)/5 = 1.60$ , so  $0.5000 - 0.4452 = 0.0548$

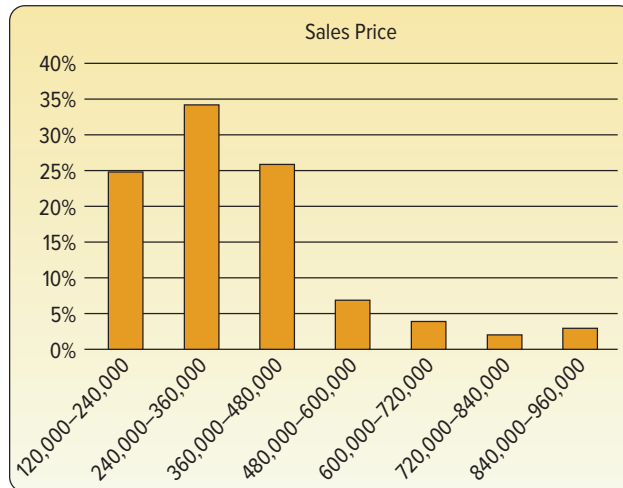
b. Let  $z = 0.67$ , so  $0.67 = (x - 52)/5$  and  $x = 55.35$ , set mileage at  $55,350$

c.  $z = (45 - 52)/5 = -1.40$ , so  $0.5000 - 0.4192 = 0.0808$

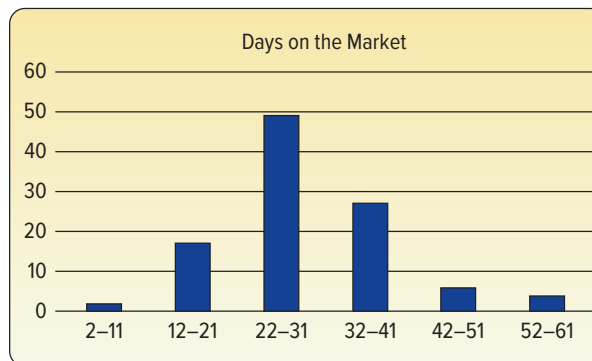
$$51. \frac{470 - \mu}{\sigma} = 0.25 \quad \frac{500 - \mu}{\sigma} = 1.28 \quad \sigma = 29,126 \text{ and } \mu = 462,718$$

53. a.  $z = (500.0 - 357.0)/160.7 = 0.89$ ;  $P(z > 0.89) = 0.5000 - 0.3133 = 0.1867$

If price was normally distributed,  $18.67\%$  or  $19.6$  homes would have sold for more than  $\$500,000$ . There are  $14$  homes or  $13.3\%$  that actually sold for more than  $\$500,000$ . If price were normally distributed,  $50\%$ , or about  $52.5$ , of the homes would sell for more than the mean ( $\$357.0$ ). The data show that  $43$  homes sold for more than the average. Conclusion: Price is probably not normally distributed.

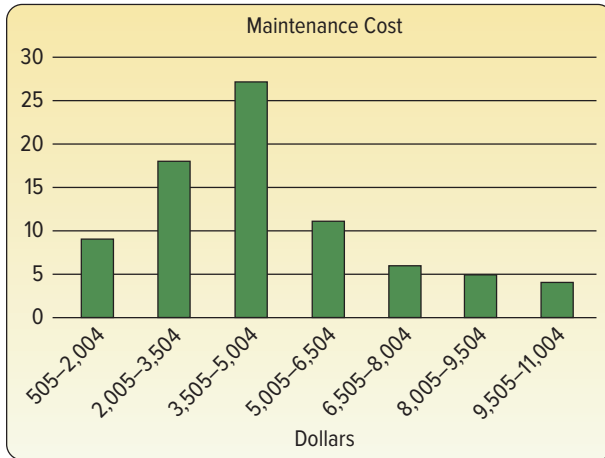


b.  $z = (24 - 30)/10 = -0.60$ ; The probability is  $0.5000 + 0.2257 = 0.7257$ . If days-on-the-market were normally distributed,  $72.57\%$  of  $105$  or  $72.6$  homes would be on the market more than  $24$  days. There were actually  $74$  homes on the market more than  $24$  days. If days-on-the-market were normally distributed,  $50\%$ , or about  $52.5$  of the homes would be on the market for more than the mean of  $30$  days. The data show that  $43$  homes were on the market longer than the average. Conclusion: Days-on-the-market could be normally distributed. Create a frequency distribution and observe its shape.

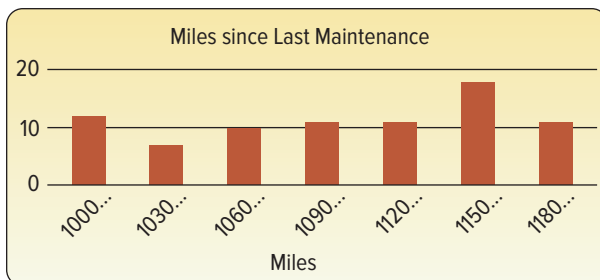


55. a.  $0.2676$ , found by  $0.5000 - 0.2324$  with  $z = (6000 - 4552)/2332 = 0.62$ ; leads to  $21.4$  buses, found by

80(0.2676). 17 buses actually had maintenance cost of more than \$6,000. So the estimate is close. The frequency distribution of maintenance cost shows a nearly normal distribution. However, it is a slightly right skewed distribution. Consequently, the estimates based on the assumption of a normal distribution may not be accurate.



b. 0.2709, found by  $0.5000 - 0.2291$  with  $z = (11,500 - 11,121)/617 = 0.61$ ; leads to 21.7 buses, found by  $80(0.2709)$ . 29 buses actually traveled more than 11,500 miles. So the estimate is not precise. The frequency distribution of miles driven since the last maintenance shows a nearly uniform distribution. Consequently, the estimates based on the assumption of a normal distribution are not accurate.



## CHAPTER 8

- 303 Louisiana, 5155 S. Main, 3501 Monroe, 2652 W. Central
  - Answers will vary.
  - 630 Dixie Hwy, 835 S. McCord Rd, 4624 Woodville Rd
  - Answers will vary.
- Bob Schmidt Chevrolet  
Great Lakes Ford Nissan  
Grogan Towne Chrysler  
Southside Lincoln  
Rouen Chrysler Jeep Eagle
  - Answers will vary.
  - York Automotive Group  
Thayer Chevrolet/Toyota  
Franklin Park Lincoln  
Mathews Ford Oregon Inc.  
Valiton Chrysler

5. a.

| Sample | Values | Sum | Mean |
|--------|--------|-----|------|
| 1      | 12, 12 | 24  | 12   |
| 2      | 12, 14 | 26  | 13   |
| 3      | 12, 16 | 28  | 14   |
| 4      | 12, 14 | 26  | 13   |
| 5      | 12, 16 | 28  | 14   |
| 6      | 14, 16 | 30  | 15   |

b.  $\mu_{\bar{x}} = (12 + 13 + 14 + 13 + 14 + 15)/6 = 13.5$

$\mu = (12 + 12 + 14 + 16)/4 = 13.5$

c. More dispersion with population data compared to the sample means. The sample means vary from 12 to 15, whereas the population varies from 12 to 16.

7. a.

| Sample | Values     | Sum | Mean  |
|--------|------------|-----|-------|
| 1      | 12, 12, 14 | 38  | 12.66 |
| 2      | 12, 12, 15 | 39  | 13.00 |
| 3      | 12, 12, 20 | 44  | 14.66 |
| 4      | 14, 15, 20 | 49  | 16.33 |
| 5      | 12, 14, 15 | 41  | 13.66 |
| 6      | 12, 14, 15 | 41  | 13.66 |
| 7      | 12, 15, 20 | 47  | 15.66 |
| 8      | 12, 15, 20 | 47  | 15.66 |
| 9      | 12, 14, 20 | 46  | 15.33 |
| 10     | 12, 14, 20 | 46  | 15.33 |

b.  $\mu_{\bar{x}} = \frac{(12.66 + \dots + 15.33 + 15.33)}{10} = 14.6$

$\mu = (12 + 12 + 14 + 15 + 20)/5 = 14.6$

c. The dispersion of the population is greater than that of the sample means. The sample means vary from 12.66 to 16.33, whereas the population varies from 12 to 20.

9. a. 20, found by  ${}_6C_3$

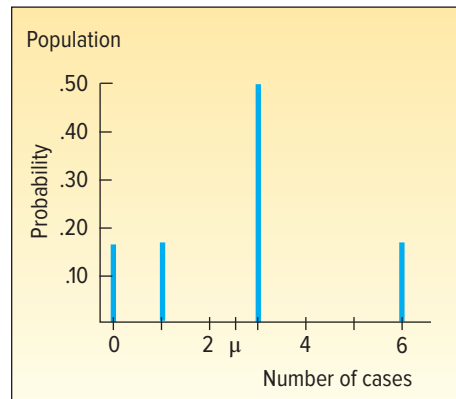
b.

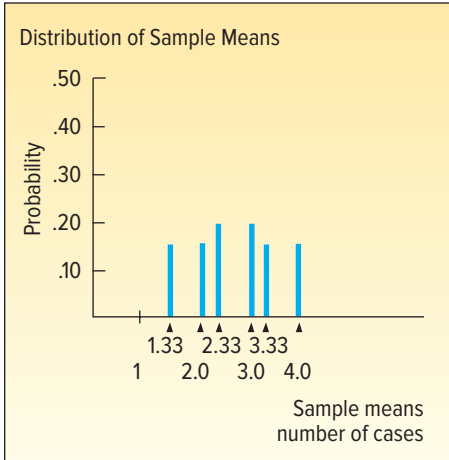
| Sample                  | Cases   | Sum | Mean |
|-------------------------|---------|-----|------|
| Ruud, Wu, Sass          | 3, 6, 3 | 12  | 4.00 |
| Ruud, Sass, Flores      | 3, 3, 3 | 9   | 3.00 |
| ⋮                       | ⋮       | ⋮   | ⋮    |
| Sass, Flores, Schueller | 3, 3, 1 | 7   | 2.33 |

c.  $\mu_{\bar{x}} = 2.67$ , found by  $\frac{53.33}{20}$

$\mu = 2.67$ , found by  $(3 + 6 + 3 + 3 + 0 + 1)/6$   
They are equal.

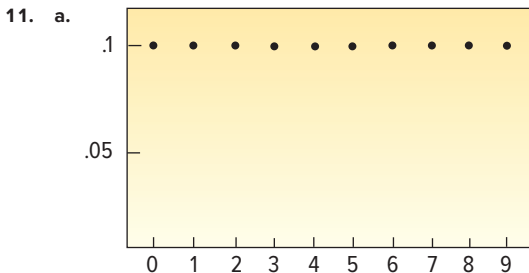
d.





| Sample Mean | Number of Means | Probability |
|-------------|-----------------|-------------|
| 1.33        | 3               | .1500       |
| 2.00        | 3               | .1500       |
| 2.33        | 4               | .2000       |
| 3.00        | 4               | .2000       |
| 3.33        | 3               | .1500       |
| 4.00        | 3               | .1500       |
|             | 20              | 1.0000      |

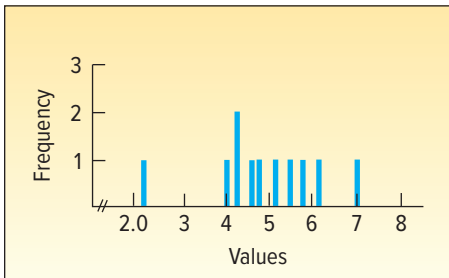
The population has more dispersion than the sample means. The sample means vary from 1.33 to 4.0. The population varies from 0 to 6.



$$\mu = \frac{0 + 1 + \dots + 9}{10} = 4.5$$

b.

| Sample | Sum | $\bar{x}$ | Sample | Sum | $\bar{x}$ |
|--------|-----|-----------|--------|-----|-----------|
| 1      | 11  | 2.2       | 6      | 20  | 4.0       |
| 2      | 31  | 6.2       | 7      | 23  | 4.6       |
| 3      | 21  | 4.2       | 8      | 29  | 5.8       |
| 4      | 24  | 4.8       | 9      | 35  | 7.0       |
| 5      | 21  | 4.2       | 10     | 27  | 5.4       |



The mean of the 10 sample means is 4.84, which is close to the population mean of 4.5. The sample means range from 2.2 to 7.0, whereas the population values range from 0 to 9. From the above graph, the sample means tend to cluster between 4 and 5.

13. a.-c. Answers will vary depending on the coins in your possession.

15. a.  $z = \frac{63 - 60}{12/\sqrt{9}} = 0.75$   
 $p = .2266$ , found by  $.5000 - .2734$

b.  $z = \frac{56 - 60}{12/\sqrt{9}} = -1.00$   
 $p = .1587$ , found by  $.5000 - .3413$

c.  $p = .6147$ , found by  $0.3413 + 0.2734$

17.  $z = \frac{1,950 - 2,200}{250/\sqrt{50}} = -7.07$   $p = 1$ , or virtually certain

19. a. Formal Man, Summit Stationers, Bootleggers, Leather Ltd., Petries

b. Answers may vary.

c. Elder-Beerman, Frederick's of Hollywood, Summit Stationers, Lion Store, Leather Ltd., Things Remembered, County Seat, Coach House Gifts, Regis Hairstylists

21. a.

| Samples | Mean | Deviation from Mean | Square of Deviation |
|---------|------|---------------------|---------------------|
| 1, 1    | 1.0  | -1.0                | 1.0                 |
| 1, 2    | 1.5  | -0.5                | 0.25                |
| 1, 3    | 2.0  | 0.0                 | 0.0                 |
| 2, 1    | 1.5  | -0.5                | 0.25                |
| 2, 2    | 2.0  | 0.0                 | 0.0                 |
| 2, 3    | 2.5  | 0.5                 | 0.25                |
| 3, 1    | 2.0  | 0.0                 | 0.0                 |
| 3, 2    | 2.5  | 0.5                 | 0.25                |
| 3, 3    | 3.0  | 1.0                 | 1.0                 |

b. Mean of sample means is  $(1.0 + 1.5 + 2.0 + \dots + 3.0)/9 = 18/9 = 2.0$ . The population mean is  $(1 + 2 + 3)/3 = 6/3 = 2$ . They are the same value.

c. Variance of sample means is  $(1.0 + 0.25 + 0.0 + \dots + 3.0)/9 = 3/9 = 1/3$ . Variance of the population values is  $(1 + 0 + 1)/3 = 2/3$ . The variance of the population is twice as large as that of the sample means.

d. Sample means follow a triangular shape, peaking at 2. The population is uniform between 1 and 3.

23. Larger samples provide narrower estimates of a population mean. So the company with 200 sampled customers can provide more precise estimates. In addition, they selected consumers who are familiar with laptop computers and may be better able to evaluate the new computer.

25. a. We selected 60, 104, 75, 72, and 48. Answers will vary.

b. We selected the third observation. So the sample consists of 75, 72, 68, 82, and 48. Answers will vary.

c. Number the first 20 motels from 00 to 19. Randomly select three numbers. Then number the last five numbers 20 to 24. Randomly select two numbers from that group.

27. a.  $(79 + 64 + 84 + 82 + 92 + 77)/6 = 79.67\%$

b. 15, found by  ${}_6C_2$

c.

| Sample | Value  | Sum | Mean    |
|--------|--------|-----|---------|
| 1      | 79, 64 | 143 | 71.5    |
| 2      | 79, 84 | 163 | 81.5    |
| ⋮      | ⋮      | ⋮   | ⋮       |
| 15     | 92, 77 | 169 | 84.5    |
|        |        |     | 1,195.0 |

- d.  $\mu_{\bar{x}} = 79.67$ , found by  $1,195/15$   
 $\mu = 79.67$ , found by  $478/6$   
 They are equal.
- e. Answers will vary. Not likely as the student is not graded on all available information. Based on these test scores, however, this student has an  $8/15$  chance of receiving a higher grade with this method than the average and a  $7/15$  chance of receiving a lower grade.

29. a. 10, found by  ${}_5C_2$

| Number of Shutdowns | Mean | Number of Shutdowns | Mean |
|---------------------|------|---------------------|------|
| 4, 3                | 3.5  | 3, 3                | 3.0  |
| 4, 5                | 4.5  | 3, 2                | 2.5  |
| 4, 3                | 3.5  | 5, 3                | 4.0  |
| 4, 2                | 3.0  | 5, 2                | 3.5  |
| 3, 5                | 4.0  | 3, 2                | 2.5  |

| Sample Mean | Frequency | Probability |
|-------------|-----------|-------------|
| 2.5         | 2         | .20         |
| 3.0         | 2         | .20         |
| 3.5         | 3         | .30         |
| 4.0         | 2         | .20         |
| 4.5         | 1         | .10         |
|             | 10        | 1.00        |

- c.  $\mu_{\bar{x}} = (3.5 + 4.5 + \dots + 2.5)/10 = 3.4$   
 $\mu = (4 + 3 + 5 + 3 + 2)/5 = 3.4$   
 The two means are equal.
- d. The population values are relatively uniform in shape. The distribution of sample means tends toward normality.
31. a. The distribution will be normal.
- b.  $\sigma_{\bar{x}} = \frac{5.5}{\sqrt{25}} = 1.1$
- c.  $z = \frac{36 - 35}{5.5/\sqrt{25}} = 0.91$   
 $p = 0.1814$ , found by  $0.5000 + 0.3186$
- d.  $z = \frac{34.5 - 35}{5.5/\sqrt{25}} = -0.45$   
 $p = 0.6736$ , found by  $0.5000 + 0.1736$
- e.  $0.4922$ , found by  $0.3186 + 0.1736$
33.  $z = \frac{\$335 - \$350}{\$45/\sqrt{40}} = -2.11$   
 $p = 0.9826$ , found by  $0.5000 + 0.4826$
35.  $z = \frac{29.3 - 29}{2.5/\sqrt{60}} = 0.93$   
 $p = 0.8238$ , found by  $0.5000 + 0.3238$
37. Between 5,954 and 6,046, found by  $6,000 \pm 1.96(150/\sqrt{40})$
39.  $z = \frac{900 - 947}{205/\sqrt{60}} = -1.78$   
 $p = 0.0375$ , found by  $0.5000 - 0.4625$
41. a. Alaska, Connecticut, Georgia, Kansas, Nebraska, South Carolina, Virginia, Utah  
 b. Arizona, Florida, Iowa, Massachusetts, Nebraska, North Carolina, Rhode Island, Vermont
43. a.  $z = \frac{600 - 510}{14.28/\sqrt{10}} = 19.9$ ,  $P = 0.00$ , or virtually never  
 b.  $z = \frac{500 - 510}{14.28/\sqrt{10}} = -2.21$   
 $p = 0.4864 + 0.5000 = 0.9864$   
 c.  $z = \frac{500 - 510}{14.28/\sqrt{10}} = -2.21$   
 $p = 0.5000 - 0.4864 = 0.0136$

45. a.  $\sigma_{\bar{x}} = \frac{2.1}{\sqrt{81}} = 0.23$   
 b.  $z = \frac{7.0 - 6.5}{2.1/\sqrt{81}} = 2.14$ ,  $z = \frac{6.0 - 6.5}{2.1/\sqrt{81}} = -2.14$ ,  
 $p = .4838 + .4838 = .9676$   
 c.  $z = \frac{6.75 - 6.5}{2.1/\sqrt{81}} = 1.07$ ,  $z = \frac{6.25 - 6.5}{2.1/\sqrt{81}} = -1.07$ ,  
 $p = .3577 + .3577 = .7154$   
 d. .0162, found by  $.5000 - .4838$
47. Mean 2016 attendance is 2.439 million. Likelihood of a sample mean this large or larger is 0.5359, found by  $0.5000 + 0.0359$ , where  $z = \frac{2.439 - 2.45}{\frac{0.71}{\sqrt{30}}} = -0.09$ .

## CHAPTER 9

1. 51.314 and 58.686, found by  $55 \pm 2.58(10/\sqrt{49})$
3. a. 1.581, found by  $\sigma_{\bar{x}} = 25/\sqrt{250}$   
 b. The population is normally distributed and the population variance is known. In addition, the central limit theorem says that the sampling distribution of sample means will be normally distributed.  
 c. 16.901 and 23.099, found by  $20 \pm 3.099$
5. a. \$20. It is our best estimate of the population mean.  
 b. \$18.60 and \$21.40, found by  $\$20 \pm 1.96(\$5/\sqrt{49})$ . About 95% of the intervals similarly constructed will include the population mean.
7. a. 8.60 gallons  
 b. 7.83 and 9.37, found by  $8.60 \pm 2.58(2.30/\sqrt{60})$   
 c. If 100 such intervals were determined, the population mean would be included in about 99 intervals.
9. a. 2.201  
 b. 1.729  
 c. 3.499
11. a. The population mean is unknown, but the best estimate is 20, the sample mean.  
 b. Use the  $t$  distribution since the standard deviation is unknown. However, assume the population is normally distributed.  
 c. 2.093  
 d. Between 19.06 and 20.94, found by  $20 \pm 2.093(2/\sqrt{20})$   
 e. Neither value is reasonable because they are not inside the interval.
13. Between 95.39 and 101.81, found by  $98.6 \pm 1.833(5.54/\sqrt{10})$
15. a. 0.8, found by  $80/100$   
 b. Between 0.72 and 0.88, found by  
 $0.8 \pm 1.96 \left( \sqrt{\frac{0.8(1 - 0.8)}{100}} \right)$   
 c. We are reasonably sure the population proportion is between 72% and 88%.
17. a. 0.625, found by  $250/400$   
 b. Between 0.563 and 0.687, found by  
 $0.625 \pm 2.58 \left( \sqrt{\frac{0.625(1 - 0.625)}{400}} \right)$   
 c. We are reasonably sure the population proportion is between 56% and 69%. Because the estimated population proportion is more than 50%, the results indicate that Fox TV should schedule the new comedy show.
19. 97, found by  $n = \left( \frac{1.96 \times 10}{2} \right)^2 = 96.04$
21. 196, found by  $n = 0.15(0.85) \left( \frac{1.96}{0.05} \right)^2 = 195.9216$
23. 554, found by  $n = \left( \frac{1.96 \times 3}{0.25} \right)^2 = 553.19$

25. a. 577, found by  $n = 0.60(0.40) \left( \frac{1.96}{0.04} \right)^2 = 576.24$   
 b. 601, found by  $n = 0.50(0.50) \left( \frac{1.96}{0.04} \right)^2 = 600.25$
27. 6.13 years to 6.87 years, found by  $6.5 \pm 1.989(1.7/\sqrt{85})$
29. a. Between \$864.82 and 903.18, found by  $884 \pm 2.426 \left( \frac{50}{\sqrt{40}} \right)$   
 b. \$950 is not reasonable because it is outside the confidence interval.
31. a. The population mean is unknown.  
 b. Between 7.50 and 9.14, found by  $8.32 \pm 1.685(3.07/\sqrt{40})$   
 c. 10 is not reasonable because it is outside the confidence interval.
33. a. 65.49 up to 71.71 hours, found by  $68.6 \pm 2.680(8.2/\sqrt{50})$   
 b. The value suggested by the NCAA is included in the confidence interval. Therefore, it is reasonable.  
 c. Changing the confidence interval to 95 would reduce the width of the interval. The value of 2.680 would change to 2.010.
35. 61.47, rounded to 62. Found by solving for  $n$  in the equation:  $1.96(16/\sqrt{n}) = 4$
37. \$55,461.23 up to \$57,769.43, found by  $55,051 \pm 1.711 \left( \frac{7,568}{\sqrt{25}} \right)$ .  
 55,000 is reasonable because it is inside the confidence interval.
39. a. 82.58, found by  $991/12$   
 b. Between 80.54 and \$84.62, found by  $82.58 \pm 1.796 \left( \frac{3.94}{\sqrt{12}} \right)$   
 c. 80 hours per week is not reasonable because it is outside the confidence interval.
41. a. 89.4667, found by  $1,342/15$   
 b. Between 84.99 and 93.94, found by  $89.4667 \pm 2.145(8.08/\sqrt{15})$   
 c. Yes, because even the lower limit of the confidence interval is above 80.
43. The confidence interval is between 0.011 and 0.059, found by  $0.035 \pm 2.576 \left( \sqrt{\frac{0.035(1 - 0.035)}{400}} \right)$ . It would not be reasonable to conclude that fewer than 5% of the employees are now failing the test because 0.05 is inside the confidence interval.
45. Between 0.648 and 0.752, found by  $0.70 \pm 2.576 \left( \sqrt{\frac{0.70(1 - 0.70)}{500}} \right) \left( \sqrt{\frac{20,000 - 500}{20,000 - 1}} \right)$ .  
 Yes, because even the lower limit of the confidence interval is above 0.500.
47. 369, found by  $n = 0.60(1 - 0.60)(1.96/0.05)^2$
49. 97, found by  $[(1.96 \times 500)/100]^2$
51. a. Between 7,849 and 8,151, found by  $8,000 \pm 2.756(300/\sqrt{30})$   
 b. 554, found by  $n = \left( \frac{(1.96)(300)}{25} \right)^2$
53. a. Between 75.44 and 80.56, found by  $78 \pm 2.010(9/\sqrt{50})$   
 b. 220, found by  $n = \left( \frac{(1.645)(9)}{1.0} \right)^2$
55. a. 4, found by  $24/\sqrt{36}$   
 b. Between \$641.88 and \$658.12, found by  $650 \pm 2.030 \left( \frac{24}{\sqrt{36}} \right)$   
 c. 23, found by  $n = [(1.96 \times 24)/10]^2 = 22.13$
57. a. 708.13, rounded up to 709, found by  $0.21(1 - 0.21)(1.96/0.03)^2$   
 b. 1,068, found by  $0.50(0.50)(1.96/0.03)^2$

59. a. Between 0.156 and 0.184, found by  $0.17 \pm 1.96 \sqrt{\frac{(0.17)(1 - 0.17)}{2,700}}$   
 b. Yes, because 18% is inside the confidence interval.  
 c. 21,682, found by  $0.17(1 - 0.17)[1.96/0.005]^2$
61. Between 12.69 and 14.11, found by  $13.4 \pm 1.96(6.8/\sqrt{352})$
63. a. Answers will vary.  
 b. Answers will vary.  
 c. Answers will vary.  
 d. Answers may vary.  
 e. Select a different sample of 20 homes and compute a confidence interval using the new sample. There is a 5% probability that a sample mean will be more than 1.96 standard errors from the mean. If this happens, the confidence interval will not include the population mean.
65. a. Between \$4,033.1476 and \$5,070.6274, found by  $4,551.8875 \pm 518.7399$   
 b. Between 71,040.0894 and 84,877.1106, found by  $77,958.6000 \pm 6,918.5106$   
 c. In general, the confidence intervals indicate that the average maintenance cost and the average odometer reading suggest an aging bus fleet.

## CHAPTER 10

1. a. Two-tailed  
 b. Reject  $H_0$  when  $z$  does not fall in the region between  $-1.96$  and  $1.96$ .  
 c.  $-1.2$ , found by  $z = (49 - 50)/(5/\sqrt{36}) = -1.2$   
 d. Fail to reject  $H_0$ .  
 e.  $p = .2302$ , found by  $2(.5000 - .3849)$ . A 23.02% chance of finding a  $z$  value this large when  $H_0$  is true.
3. a. One-tailed  
 b. Reject  $H_0$  when  $z > 1.65$ .  
 c.  $1.2$ , found by  $z = (21 - 20)/(5/\sqrt{36})$   
 d. Fail to reject  $H_0$  at the .05 significance level.  
 e.  $p = .1151$ , found by  $.5000 - .3849$ . An 11.51% chance of finding a  $z$  value this large or larger.
5. a.  $H_0: \mu = 60,000$      $H_1: \mu \neq 60,000$   
 b. Reject  $H_0$  if  $z < -1.96$  or  $z > 1.96$ .  
 c.  $-0.69$ , found by:  

$$z = \frac{59,500 - 60,000}{(5,000/\sqrt{48})}$$
- d. Do not reject  $H_0$ .  
 e.  $p = .4902$ , found by  $2(.5000 - .2549)$ . Crosset's experience is not different from that claimed by the manufacturer. If  $H_0$  is true, the probability of finding a value more extreme than this is .4902.
7. a.  $H_0: \mu \geq 6.8$      $H_1: \mu < 6.8$   
 b. Reject  $H_0$  if  $z < -1.65$   
 c.  $z = \frac{6.2 - 6.8}{1.8/\sqrt{36}} = -2.0$   
 d.  $H_0$  is rejected.  
 e.  $p = 0.0228$ . The mean number of DVDs watched is less than 6.8 per month. If  $H_0$  is true, you will get a statistic this small less than one time out of 40 tests.
9. a. Reject  $H_0$  when  $t < 1.833$ .  
 b.  $t = \frac{12 - 10}{(3/\sqrt{10})} = 2.108$   
 c. Reject  $H_0$ . The mean is greater than 10.
11.  $H_0: \mu \leq 40$      $H_1: \mu > 40$   
 Reject  $H_0$  if  $t > 1.703$ .  

$$t = \frac{42 - 40}{(2.1/\sqrt{28})} = 5.040$$

Reject  $H_0$  and conclude that the mean number of calls is greater than 40 per week.

13.  $H_0: \mu \leq 50,000$   $H_1: \mu > 50,000$   
Reject  $H_0$  if  $t > 1.833$ .

$$t = \frac{(60,000 - 50,000)}{(10,000/\sqrt{10})} = 3.16$$

Reject  $H_0$  and conclude that the mean income in Wilmington is greater than \$50,000.

15. a. Reject  $H_0$  if  $t < -3.747$ .

b.  $\bar{x} = 17$  and  $s = \sqrt{\frac{50}{5-1}} = 3.536$

$$t = \frac{17 - 20}{(3.536/\sqrt{5})} = -1.90$$

c. Do not reject  $H_0$ . We cannot conclude the population mean is less than 20.

d. Between .05 and .10, about .065

17.  $H_0: \mu \leq 1.4$   $H_1: \mu > 1.4$

Reject  $H_0$  if  $t > 2.821$ .

$$t = \frac{1.6 - 1.4}{0.216/\sqrt{10}} = 2.93$$

Reject  $H_0$  and conclude that water consumption has increased. The  $p$ -value is between 0.01 and 0.005. There is a slight probability (between one chance in 100 and one chance in 200) this increase could have arisen by chance.

19.  $H_0: \mu \leq 67$   $H_1: \mu > 67$

Reject  $H_0$  if  $t > 1.796$ .

$$t = \frac{(82.5 - 67)}{(59.5/\sqrt{12})} = 0.902$$

Fail to reject  $H_0$  and conclude that the mean number of text messages is not greater than 67. The  $p$ -value is greater than 0.05. There is a good probability (about 18%) this could happen by chance.

21.  $H_0: \mu = \$45,000$   $H_1: \mu \neq \$45,000$

Reject  $H_0$  if  $z < -1.65$  or  $z > 1.65$ .

$$z = \frac{\$45,500 - \$45,000}{\$3,000/\sqrt{120}} = 1.83$$

Reject  $H_0$ . We can conclude that the mean salary is not \$45,000.  $p$ -value = 0.0672, found by  $2(0.5000 - 0.4664)$

23.  $H_0: \mu \geq 10$   $H_1: \mu < 10$

Reject  $H_0$  if  $z < -1.65$ .

$$z = \frac{9.0 - 10.0}{2.8/\sqrt{50}} = -2.53$$

Reject  $H_0$ . The mean weight loss is less than 10 pounds.  $p$ -value =  $0.5000 - 0.4943 = 0.0057$

25.  $H_0: \mu \geq 7.0$   $H_1: \mu < 7.0$

Assuming a 5% significance level, reject  $H_0$  if  $t < -1.677$ .

$$t = \frac{6.8 - 7.0}{0.9/\sqrt{50}} = -1.57$$

Do not reject  $H_0$ . West Virginia students are not sleeping less than 6 hours.  $p$ -value is between .05 and .10.

27.  $H_0: \mu \geq 3.13$   $H_1: \mu < 3.13$

Reject  $H_0$  if  $t < -1.711$ .

$$t = \frac{2.86 - 3.13}{1.20/\sqrt{25}} = -1.13$$

We fail to reject  $H_0$  and conclude that the mean number of residents is not necessarily less than 3.13.

29.  $H_0: \mu \leq \$6,658$   $H_1: \mu > \$6,658$

Reject  $H_0$  if  $t > 1.796$ .

$$\bar{x} = \frac{85,963}{12} = 7,163.58 \quad s = \sqrt{\frac{9,768,674.92}{12-1}} = 942.37$$

$$t = \frac{7163.58 - 6,658}{942.37/\sqrt{12}} = 1.858$$

Reject  $H_0$ . The mean interest paid is greater than \$6,658.

31.  $H_0: \mu = 3.1$   $H_1: \mu \neq 3.1$  Assume a normal population.

Reject  $H_0$  if  $t < -2.201$  or  $t > 2.201$ .

$$\bar{x} = \frac{41.1}{12} = 3.425$$

$$s = \sqrt{\frac{4.0625}{12-1}} = .6077$$

$$t = \frac{3.425 - 3.1}{.6077/\sqrt{12}} = 1.853$$

Do not reject  $H_0$ . Cannot show a difference between senior citizens and the national average.  $p$ -value is about 0.09.

33.  $H_0: \mu \geq 6.5$   $H_1: \mu < 6.5$  Assume a normal population.

Reject  $H_0$  if  $t < -2.718$ .

$$\bar{x} = 5.1667 \quad s = 3.1575$$

$$t = \frac{5.1667 - 6.5}{3.1575/\sqrt{12}} = -1.463$$

Do not reject  $H_0$ . The  $p$ -value is greater than 0.05.

35.  $H_0: \mu = 0$   $H_1: \mu \neq 0$

Reject  $H_0$  if  $t < -2.110$  or  $t > 2.110$ .

$$\bar{x} = -0.2322 \quad s = 0.3120$$

$$t = \frac{-0.2322 - 0}{0.3120/\sqrt{18}} = -3.158$$

Reject  $H_0$ . The mean gain or loss does not equal 0. The  $p$ -value is less than 0.01, but greater than 0.001.

37.  $H_0: \mu \leq 100$   $H_1: \mu > 100$  Assume a normal population.

Reject  $H_0$  if  $t > 1.761$ .

$$\bar{x} = \frac{1,641}{15} = 109.4$$

$$s = \sqrt{\frac{1,389.6}{15-1}} = 9.9628$$

$$t = \frac{109.4 - 100}{9.9628/\sqrt{15}} = 3.654$$

Reject  $H_0$ . The mean number with the scanner is greater than 100.  $p$ -value is 0.001.

39.  $H_0: \mu = 1.5$   $H_1: \mu \neq 1.5$

Reject  $H_0$  if  $t > 3.250$  or  $t < -3.250$ .

$$t = \frac{1.3 - 1.5}{0.9/\sqrt{10}} = -0.703$$

Fail to reject  $H_0$ .

41.  $H_0: \mu \geq 30$   $H_1: \mu < 30$

Reject  $H_0$  if  $t < -1.895$ .

$$\bar{x} = \frac{238.3}{8} = 29.7875 \quad s = \sqrt{\frac{5.889}{8-1}} = 0.9172$$

$$t = \frac{29.7875 - 30}{0.9172/\sqrt{8}} = -0.655$$

Do not reject  $H_0$ . The cost is not less than \$30,000.

43. a.  $9.00 \pm 1.645(1/\sqrt{36}) = 9.00 \pm 0.274$ .

So the limits are 8.726 and 9.274.

b.  $z = \frac{8.726 - 8.6}{1/\sqrt{36}} = 0.756$

$$P(z < 0.756) = 0.5000 + 0.2764 = .7764$$

c.  $z = \frac{9.274 - 9.6}{1/\sqrt{36}} = -1.956$

$$P(z > -1.96) = 0.4750 + 0.5000 = .9750$$

45. a.  $H_0: \mu = 100$   $H_1: \mu \neq 100$

Reject  $H_0$  if  $t$  is not between  $-2.045$  and  $2.045$ .

$$t = \frac{121.12 - 100}{39.66/\sqrt{30}} = 2.92$$

Reject the null. The mean salary is probably not \$100.0 million.

- b.  $H_0: \mu \leq 2,000,000$   $H_1: \mu > 2,000,000$

Reject  $H_0$  if  $t$  is  $> 1.699$ .

$$t = \frac{2,438,636 - 2,000,000}{617,670/\sqrt{30}} = 3.89$$

Reject the null. The mean attendance was more than 2,000,000.



## CHAPTER 11

1. a. Two-tailed test  
 b. Reject  $H_0$  if  $z < -2.05$  or  $z > 2.05$ .  
 c.  $z = \frac{102 - 99}{\sqrt{\frac{5^2}{40} + \frac{6^2}{50}}} = 2.59$   
 d. Reject  $H_0$ .  
 e.  $p$ -value = .0096, found by  $2(.5000 - .4952)$
3. **Step 1**  $H_0: \mu_1 \geq \mu_2$      $H_1: \mu_1 < \mu_2$   
**Step 2** The .05 significance level was chosen.  
**Step 3** Reject  $H_0$  if  $z < -1.65$ .  
**Step 4**  $-0.94$ , found by:

$$z = \frac{7.6 - 8.1}{\sqrt{\frac{(2.3)^2}{40} + \frac{(2.9)^2}{55}}} = -0.94$$

**Step 5** Fail to reject  $H_0$ .

**Step 6** Babies using the Gibbs brand did not gain less weight.  $p$ -value = .1736, found by  $.5000 - .3264$

5. **Step 1**  $H_0: \mu_{\text{married}} = \mu_{\text{unmarried}}$      $H_1: \mu_{\text{married}} \neq \mu_{\text{unmarried}}$   
**Step 2** The 0.05 significance level was chosen.  
**Step 3** Use a  $z$ -statistic, as both population standard deviations are known.  
**Step 4** If  $z < -1.960$  or  $z > 1.960$ , reject  $H_0$ .  
**Step 5**  $z = \frac{3.0 - 3.4}{\sqrt{\frac{(1.2)^2}{45} + \frac{(1.1)^2}{39}}} = -1.59$

Fail to reject the null.

**Step 6** It is reasonable to conclude that the time that married and unmarried women spend each week is not significantly different. The  $p$ -value is greater than 0.05. The difference of 0.4 hours per week could be explained by sampling error.

7. a. Reject  $H_0$  if  $t > 2.120$  or  $t < -2.120$ .  $df = 10 + 8 - 2 = 16$   
 b.  $s_p^2 = \frac{(10 - 1)(4)^2 + (8 - 1)(5)^2}{10 + 8 - 2} = 19.9375$   
 c.  $t = \frac{23 - 26}{\sqrt{19.9375 \left( \frac{1}{10} + \frac{1}{8} \right)}} = -1.416$   
 d. Do not reject  $H_0$ .  
 e.  $p$ -value is greater than .10 and less than .20.

9. **Step 1**  $H_0: \mu_{\text{Pitchers}} = \mu_{\text{Position Players}}$   
 $H_1: \mu_{\text{Pitchers}} \neq \mu_{\text{Position Players}}$   
**Step 2** The 0.01 significance level was chosen.  
**Step 3** Use a  $t$ -statistic, assuming a pooled variance with the standard deviation unknown.  
**Step 4**  $df = 12 + 13 - 2 = 23$ . Reject  $H_0$  if  $t$  is not between  $-2.807$  and  $2.807$ .  
 $s_p^2 = \frac{(12 - 1)(8.597)^2 + (13 - 1)(8.578)^2}{12 + 13 - 2} = 73.738$   
 $t = \frac{6.091 - 10.684}{\sqrt{73.738 \left( \frac{1}{12} + \frac{1}{13} \right)}} = -1.336$

**Step 5** Do not reject  $H_0$ .

**Step 6** There is no difference in the mean salaries of pitchers and position players.

11. **Step 1**  $H_0: \mu_s \leq \mu_g$      $H_1: \mu_s > \mu_g$   
**Step 2** The .10 significance level was chosen.  
**Step 3**  $df = 6 + 7 - 2 = 11$   
 Reject  $H_0$  if  $t > 1.363$ .  
**Step 4**  $s_p^2 = \frac{(6 - 1)(12.2)^2 + (7 - 1)(15.8)^2}{6 + 7 - 2} = 203.82$   
 $t = \frac{142.5 - 130.3}{\sqrt{203.82 \left( \frac{1}{6} + \frac{1}{7} \right)}} = 1.536$   
**Step 5** Reject  $H_0$ .

**Step 6** The mean daily expenses are greater for the sales staff. The  $p$ -value is between .05 and .10.

13. Reject  $H_0$  if  $t > 2.353$ .  
 $\bar{d} = \frac{12}{4} = 3.00$      $s_d = \sqrt{\frac{2}{3}} = 0.816$   
 $t = \frac{3.00}{0.816/\sqrt{4}} = 7.35$   
 Reject  $H_0$ . There are more defective parts produced on the day shift.  
 $p$ -value is less than .005 but greater than .0005.
15.  $H_0: \mu_d \leq 0$      $H_1: \mu_d > 0$   
 $\bar{d} = 25.917$   
 $s_d = 40.791$   
 Reject  $H_0$  if  $t > 1.796$ .

$$t = \frac{25.917}{40.791/\sqrt{12}} = 2.20$$

Reject  $H_0$ . The incentive plan resulted in an increase in daily income. The  $p$ -value is about .025.

17.  $H_0: \mu_M = \mu_W$      $H_1: \mu_M \neq \mu_W$   
 Reject  $H_0$  if  $t < -2.645$  or  $t > 2.645$  ( $df = 35 + 40 - 2$ ).  
 $s_p^2 = \frac{(35 - 1)(4.48)^2 + (40 - 1)(3.86)^2}{35 + 40 - 2} = 17.3079$   
 $t = \frac{24.51 - 22.69}{\sqrt{17.3079 \left( \frac{1}{35} + \frac{1}{40} \right)}} = 1.890$

Do not reject  $H_0$ . There is no difference in the number of times men and women buy take-out dinner in a month. The  $p$ -value is between .05 and .10.

19.  $H_0: \mu_1 = \mu_2$      $H_1: \mu_1 \neq \mu_2$   
 Reject  $H_0$  if  $z < -1.96$  or  $z > 1.96$ .  
 $z = \frac{4.77 - 5.02}{\sqrt{\frac{(1.05)^2}{40} + \frac{(1.23)^2}{50}}} = -1.04$

$H_0$  is not rejected. There is no difference in the mean number of calls.  $p$ -value =  $2(.5000 - .3508) = .2984$ .

21.  $H_0: \mu_B \leq \mu_A$      $H_1: \mu_B > \mu_A$   
 Reject  $H_0$  if  $t > 1.668$ .  
 $t = \frac{\$61,000 - \$57,000}{\sqrt{\frac{(\$7,100)^2}{30} + \frac{(\$9,200)^2}{40}}} = \frac{\$4,000.00}{\$1,948.42} = 2.05$

Reject  $H_0$ . The mean income is larger for Plan B. The  $p$ -value =  $.5000 - .4798 = .0202$ .

23. Assume equal population standard deviations.  
 $H_0: \mu_n = \mu_s$      $H_1: \mu_n \neq \mu_s$   
 Reject  $H_0$  if  $t < -2.086$  or  $t > 2.086$ .  
 $s_p^2 = \frac{(10 - 1)(10.5)^2 + (12 - 1)(14.25)^2}{10 + 12 - 2} = 161.2969$   
 $t = \frac{83.55 - 78.8}{\sqrt{161.2969 \left( \frac{1}{10} + \frac{1}{12} \right)}} = 0.874$

$p$ -value  $> .10$ . Do not reject  $H_0$ . There is no difference in the mean number of hamburgers sold at the two locations.

25.  $H_0: \mu_1 \leq \mu_2$      $H_1: \mu_1 > \mu_2$     Reject  $H_0$  if  $t > 2.650$ .  
 $\bar{x}_1 = 125.125$      $s_1 = 15.094$   
 $\bar{x}_2 = 117.714$      $s_2 = 19.914$   
 $s_p^2 = \frac{(8 - 1)(15.094)^2 + (7 - 1)(19.914)^2}{8 + 7 - 2} = 305.708$   
 $t = \frac{125.125 - 117.714}{\sqrt{305.708 \left( \frac{1}{8} + \frac{1}{7} \right)}} = 0.819$

$H_0$  is not rejected. There is no difference in the mean number sold at the regular price and the mean number sold at the reduced price.

27.  $H_0: \mu_d \leq 0$      $H_1: \mu_d > 0$     Reject  $H_0$  if  $t > 1.895$ .  
 $\bar{d} = 1.75$      $s_d = 2.9155$   
 $t = \frac{1.75}{2.9155/\sqrt{8}} = 1.698$

Do not reject  $H_0$ . There is no difference in the mean number of absences. The  $p$ -value is greater than .05 but less than .10.

29.  $H_0: \mu_1 = \mu_2$      $H_1: \mu_1 \neq \mu_2$   
 Reject  $H_0$  if  $t < -2.024$  or  $t > 2.204$ .

$$s_p^2 = \frac{(15-1)(40)^2 + (25-1)(30)^2}{15+25-2} = 1,157.89$$

$$t = \frac{150-180}{\sqrt{1,157.89\left(\frac{1}{15} + \frac{1}{25}\right)}} = -2.699$$

Reject the null hypothesis. The population means are different.

31.  $H_0: \mu_d \leq 0$      $H_1: \mu_d > 0$   
 Reject  $H_0$  if  $t > 1.895$ .  
 $\bar{d} = 3.11$      $s_d = 2.91$   
 $t = \frac{3.11}{2.91/\sqrt{8}} = 3.02$

Reject  $H_0$ . The mean is lower.

33.  $H_0: \mu_O = \mu_R$      $H_1: \mu_O \neq \mu_R$   
 $df = 25 + 28 - 2 = 51$   
 Reject  $H_0$  if  $t < -2.008$  or  $t > 2.008$ .  
 $\bar{x}_O = 86.24$ ,  $s_O = 23.43$   
 $\bar{x}_R = 92.04$ ,  $s_R = 24.12$

$$s_p^2 = \frac{(25-1)(23.43)^2 + (28-1)(24.12)^2}{25+28-2} = 566.335$$

$$t = \frac{86.24 - 92.04}{\sqrt{566.335\left(\frac{1}{25} + \frac{1}{28}\right)}} = -0.886$$

Do not reject  $H_0$ . There is no difference in the mean number of cars in the two lots.

35.  $H_0: \mu_d \leq 0$      $H_1: \mu_d > 0$     Reject  $H_0$  if  $t > 1.711$ .  
 $\bar{d} = 2.8$      $s_d = 6.59$   
 $t = \frac{2.8}{6.59/\sqrt{25}} = 2.124$

Reject  $H_0$ . There are on average more cars in the US 17 lot.

37. a. Using statistical software, the result is that we fail to reject the null hypothesis that the mean prices of homes with and without pools are equal. Assuming equal population variances, the  $p$ -value is 0.4908.  
 b. Using statistical software, the result is that we reject the null hypothesis that the mean prices of homes with and without garages are equal. There is a large difference in mean prices between homes with and without garages. Assuming equal population variances, the  $p$ -value is less than 0.0001.  
 c. Using statistical software, the result is that we fail to reject the null hypothesis that the mean prices of homes are equal with mortgages in default and not in default. Assuming equal population variances, the  $p$ -value is 0.6980.  
 39. Using statistical software, the result is that we reject the null hypothesis that the mean maintenance cost of buses powered by diesel and gasoline engines is the same. Assuming equal population variances, the  $p$ -value is less than 0.0001.

## CHAPTER 12

1. a. 9.01, from Appendix B.6  
 3. Reject  $H_0$  if  $F > 10.5$ , where degrees of freedom are 7 in the numerator and 5 in the denominator. Computed  $F = 2.04$ , found by:

$$F = \frac{s_1^2}{s_2^2} = \frac{(10)^2}{(7)^2} = 2.04$$

Do not reject  $H_0$ . There is no difference in the variations of the two populations.

5.  $H_0: \sigma_1^2 = \sigma_2^2$      $H_1: \sigma_1^2 \neq \sigma_2^2$   
 Reject  $H_0$  where  $F > 3.10$ . (3.10 is about halfway between 3.14 and 3.07.) Computed  $F = 1.44$ , found by:

$$F = \frac{(12)^2}{(10)^2} = 1.44$$

Do not reject  $H_0$ . There is no difference in the variations of the two populations.

7. a.  $H_0: \mu_1 = \mu_2 = \mu_3$ ;  $H_1$ : Treatment means are not all the same.  
 b. Reject  $H_0$  if  $F > 4.26$ .

c & d.

| Source    | SS    | df | MS    | F     |
|-----------|-------|----|-------|-------|
| Treatment | 62.17 | 2  | 31.08 | 21.94 |
| Error     | 12.75 | 9  | 1.42  |       |
| Total     | 74.92 | 11 |       |       |

e. Reject  $H_0$ . The treatment means are not all the same.

9.  $H_0: \mu_1 = \mu_2 = \mu_3$ ;  $H_1$ : Treatment means are not all the same. Reject  $H_0$  if  $F > 4.26$ .

| Source    | SS     | df | MS     | F     |
|-----------|--------|----|--------|-------|
| Treatment | 276.50 | 2  | 138.25 | 14.18 |
| Error     | 87.75  | 9  | 9.75   |       |

Reject  $H_0$ . The treatment means are not all the same.

11. a.  $H_0: \mu_1 = \mu_2 = \mu_3$ ;  $H_1$ : Not all means are the same.  
 b. Reject  $H_0$  if  $F > 4.26$ .  
 c. SST = 107.20, SSE = 9.47, SS total = 116.67  
 d.

| Source    | SS     | df | MS     | F     |
|-----------|--------|----|--------|-------|
| Treatment | 107.20 | 2  | 53.600 | 50.96 |
| Error     | 9.47   | 9  | 1.052  |       |
| Total     | 116.67 | 11 |        |       |

e. Since  $50.96 > 4.26$ ,  $H_0$  is rejected. At least one of the means differs.

f.  $(\bar{x}_1 - \bar{x}_2) \pm t\sqrt{MSE(1/n_1 + 1/n_2)}$   
 $= (9.667 - 2.20) \pm 2.262\sqrt{1.052(1/3 + 1/5)}$   
 $= 7.467 \pm 1.69$   
 $= [5.777, 9.157]$

Yes, we can conclude that treatments 1 and 2 have different means.

13.  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ ;  $H_1$ : Not all means are equal.  
 $H_0$  is rejected if  $F > 3.71$ .

| Source    | SS    | df | MS    | F    |
|-----------|-------|----|-------|------|
| Treatment | 32.33 | 3  | 10.77 | 2.36 |
| Error     | 45.67 | 10 | 4.567 |      |
| Total     | 78.00 | 13 |       |      |

Because 2.36 is less than 3.71,  $H_0$  is not rejected. There is no difference in the mean number of weeks.

15.  $H_0: \sigma_1^2 \leq \sigma_2^2$ ;  $H_1: \sigma_1^2 > \sigma_2^2$ .  $df_1 = 21 - 1 = 20$ ;  
 $df_2 = 18 - 1 = 17$ .  $H_0$  is rejected if  $F > 3.16$ .

$$F = \frac{(45,600)^2}{(21,330)^2} = 4.57$$

Reject  $H_0$ . There is more variation in the selling price of ocean-front homes.

17. Sharkey:  $n = 7$   $s_s = 14.79$   
 White:  $n = 8$   $s_w = 22.95$   
 $H_0: \sigma_w^2 \leq \sigma_s^2$ ;  $H_1: \sigma_w^2 > \sigma_s^2$ ;  $df_s = 7 - 1 = 6$ ;  
 $df_w = 8 - 1 = 7$ . Reject  $H_0$  if  $F > 8.26$ .  

$$F = \frac{(22.95)^2}{(14.79)^2} = 2.41$$

Cannot reject  $H_0$ . There is no difference in the variation of the monthly sales.

19. a.  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$   
 $H_1$ : Treatment means are not all equal.  
 b.  $\alpha = .05$  Reject  $H_0$  if  $F > 3.10$ .

| Source    | SS  | df            | MS     | F    |
|-----------|-----|---------------|--------|------|
| Treatment | 50  | $4 - 1 = 3$   | $50/3$ | 1.67 |
| Error     | 200 | $24 - 4 = 20$ | 10     |      |
| Total     | 250 | $24 - 1 = 23$ |        |      |

d. Do not reject  $H_0$ .

21.  $H_0: \mu_1 = \mu_2 = \mu_3$ ;  $H_1$ : Not all treatment means are equal.  
 $H_0$  is rejected if  $F > 3.89$ .

| Source    | SS    | df | MS     | F     |
|-----------|-------|----|--------|-------|
| Treatment | 63.33 | 2  | 31.667 | 13.38 |
| Error     | 28.40 | 12 | 2.367  |       |
| Total     | 91.73 | 14 |        |       |

$H_0$  is rejected. There is a difference in the treatment means.

23.  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ ;  $H_1$ : Not all means are equal.  
 $H_0$  is rejected if  $F > 3.10$ .

| Source | SS     | df | MS    | F    |
|--------|--------|----|-------|------|
| Factor | 87.79  | 3  | 29.26 | 9.12 |
| Error  | 64.17  | 20 | 3.21  |      |
| Total  | 151.96 | 23 |       |      |

Because the computed  $F$  of 9.12  $>$  3.10, the null hypothesis of no difference is rejected at the .05 level.

25. a.  $H_0: \mu_1 = \mu_2$ ;  $H_1: \mu_1 \neq \mu_2$ . Critical value of  $F = 4.75$ .

| Source    | SS     | df | MS     | F     |
|-----------|--------|----|--------|-------|
| Treatment | 219.43 | 1  | 219.43 | 23.10 |
| Error     | 114.00 | 12 | 9.5    |       |
| Total     | 333.43 | 13 |        |       |

b.  $t = \frac{37 - 45}{\sqrt{9.5 \left( \frac{1}{6} + \frac{1}{8} \right)}} = -4.806$

$t^2 = F$ . That is,  $(-4.806)^2 \approx 23.10$  (actually 23.098, difference due to rounding). The  $p$ -value for this statistic is 0.0004 as well. Reject  $H_0$  in favor of the alternative.

c.  $H_0$  is rejected. There is a difference in the mean scores.

27. The null hypothesis is rejected because the  $F$ -statistic (8.26) is greater than the critical value (5.61) at the .01 significance level. The  $p$ -value (.0019) is also less than the significance level. The mean miles per gallon are not the same.

29.  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$   $H_1$ : At least one mean is different. Reject  $H_0$  if  $F > 2.7395$ . Since  $13.74 > 2.74$ , reject  $H_0$ . You can also see this from the  $p$ -value of 0.0001  $<$  0.05. Priority mail express is faster than all three of the other classes, and priority mail is faster than either first-class or standard. However, first-class and standard mail may be the same.

31.  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ ;  $H_1$ : The treatment means are not equal. Reject  $H_0$  if  $F > 2.37$ .

| Source    | SS      | df | MS      | F    |
|-----------|---------|----|---------|------|
| Treatment | 0.03478 | 5  | 0.00696 | 3.86 |
| Error     | 0.10439 | 58 | 0.0018  |      |
| Total     | 0.13917 | 63 |         |      |

$H_0$  is rejected. There is a difference in the mean weight of the colors.

33. a.  $H_0: \sigma_p^2 = \sigma_{np}^2$   $H_1: \sigma_p^2 \neq \sigma_{np}^2$   
 Reject  $H_0$ . The  $p$ -value is less than 0.05. There is a difference in the variance of average selling prices between houses with pools and houses without pools.  
 b.  $H_0: \sigma_g^2 = \sigma_{ng}^2$   $H_1: \sigma_g^2 \neq \sigma_{ng}^2$   
 Reject  $H_0$ . There is a difference in the variance of average selling prices between house with garages and houses without garages. The  $p$ -value is  $<$  0.0001.  
 c.  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ ;  $H_1$ : Not all treatment means are equal.

Fail to reject  $H_0$ . The  $p$ -value is much larger than 0.05. There is no statistical evidence of differences in the mean selling price between the five townships.

- d.  $H_0: \mu_c = \mu_i = \mu_m = \mu_p = \mu_r$   $H_1$ : Not all treatment means are equal.

Fail to reject  $H_0$ . The  $p$ -value is much larger than 0.05. There is no statistical evidence of differences in the mean selling price between the five agents. Is fairness of assignment based on the overall mean price, or based on the comparison of the means of the prices assigned to the agents?

While the  $p$ -value is not less than 0.05, it may indicate that the pairwise differences should be reviewed. These indicate that Marty's comparisons to the other agents are significantly different.

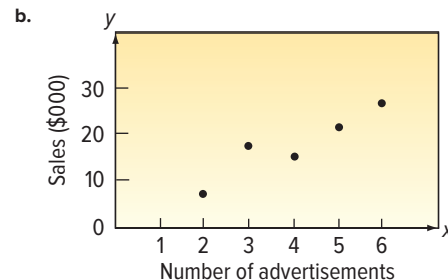
35. a.  $H_0: \mu_B = \mu_K = \mu_T$   $H_1$ : Not all treatment (manufacturer) mean maintenance costs are equal.  
 Do not reject  $H_0$ . ( $p = 0.7664$ ). The mean maintenance costs by the bus manufacturer are not different.  
 b.  $H_0: \mu_B = \mu_K = \mu_T$   $H_1$ : Not all treatments have equal mean miles since the last maintenance.  
 Do not reject  $H_0$ . The mean miles since the last maintenance by the bus manufacturer are not different.  $p$ -value = 0.4828.

## CHAPTER 13

1.  $\Sigma(x - \bar{x})(y - \bar{y}) = 10.6$ ,  $s_x = 2.7$ ,  $s_y = 1.3$

$$r = \frac{10.6}{(5 - 1)(2.709)(1.38)} = 0.75$$

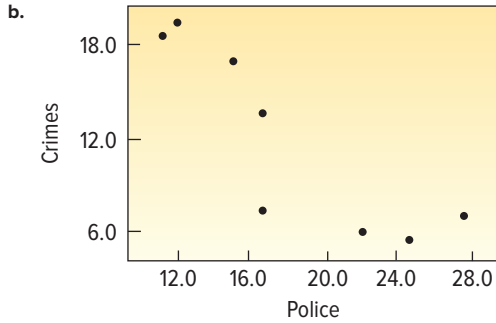
3. a. Sales



- c.  $\Sigma(x - \bar{x})(y - \bar{y}) = 36$ ,  $n = 5$ ,  $s_x = 1.5811$ ,  $s_y = 6.1237$

$$r = \frac{36}{(5 - 1)(1.5811)(6.1237)} = 0.9295$$

- d. There is a strong positive association between the variables.  
 5. a. Either variable could be independent. In the scatter plot, police is the independent variable.



c.  $n = 8$ ,  $\Sigma(x - \bar{x})(y - \bar{y}) = -231.75$ ,  
 $s_x = 5.8737$ ,  $s_y = 6.4462$

$$r = \frac{-231.75}{(8 - 1)(5.8737)(6.4462)} = -0.8744$$

d. Strong inverse relationship. As the number of police increases, the crime decreases, or as crime increases, the number of police decreases.

7. Reject  $H_0$  if  $t > 1.812$ .

$$t = \frac{.32\sqrt{12 - 2}}{\sqrt{1 - (.32)^2}} = 1.068$$

Do not reject  $H_0$ .

9.  $H_0: \rho \leq 0$ ;  $H_1: \rho > 0$ . Reject  $H_0$  if  $t > 2.552$ .  $df = 18$ .

$$t = \frac{.78\sqrt{20 - 2}}{\sqrt{1 - (.78)^2}} = 5.288$$

Reject  $H_0$ . There is a positive correlation between gallons sold and the pump price.

11.  $H_0: \rho \leq 0$   $H_1: \rho > 0$

Reject  $H_0$  if  $t > 2.650$  with  $df = 13$ .

$$t = \frac{0.667\sqrt{15 - 2}}{\sqrt{1 - 0.667^2}} = 3.228$$

Reject  $H_0$ . There is a positive correlation between the number of passengers and plane weight.

13. a.  $\hat{y} = 3.7671 + 0.3630x$

$$b = 0.7522\left(\frac{1.3038}{2.7019}\right) = 0.3630$$

$$a = 5.8 - 0.3630(5.6) = 3.7671$$

b. 6.3081, found by  $\hat{y} = 3.7671 + 0.3630(7)$

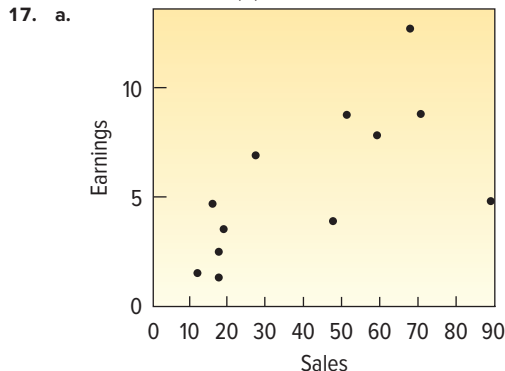
15. a.  $\Sigma(x - \bar{x})(y - \bar{y}) = 44.6$ ,  $s_x = 2.726$ ,  $s_y = 2.011$

$$r = \frac{44.6}{(10 - 1)(2.726)(2.011)} = .904$$

$$b = .904\left(\frac{2.011}{2.726}\right) = 0.667$$

$$a = 7.4 - .677(9.1) = 1.333$$

b.  $\hat{Y} = 1.333 + .667(6) = 5.335$



b.  $\Sigma(x - \bar{x})(y - \bar{y}) = 629.64$ ,  $s_x = 26.17$ ,  $s_y = 3.248$

$$r = \frac{629.64}{(12 - 1)(26.17)(3.248)} = .6734$$

c.  $b = .6734\left(\frac{3.248}{26.170}\right) = 0.0836$

$$a = \frac{64.1}{12} - 0.0836\left(\frac{501.10}{12}\right) = 1.8507$$

d.  $\hat{y} = 1.8507 + 0.0836(50.0) = 6.0307$  (\$ millions)

19. a.  $b = -.8744\left(\frac{6.4462}{5.8737}\right) = -0.9596$

$$a = \frac{95}{8} - (-0.9596)\left(\frac{146}{8}\right) = 29.3877$$

b. 10.1957, found by  $29.3877 - 0.9596(20)$

c. For each policeman added, crime goes down by almost one.

21.  $H_0: \beta \geq 0$   $H_1: \beta < 0$   $df = n - 2 = 8 - 2 = 6$   
 Reject  $H_0$  if  $t < -1.943$ .

$$t = -0.96/0.22 = -4.364$$

Reject  $H_0$  and conclude the slope is less than zero.

23.  $H_0: \beta = 0$   $H_1: \beta \neq 0$   $df = n - 2 = 12 - 2 = 10$   
 Reject  $H_0$  if  $t$  not between  $-2.228$  and  $2.228$ .

$$t = 0.08/0.03 = 2.667$$

Reject  $H_0$  and conclude the slope is different from zero.

25. The standard error of estimate is 3.378, found by  $\sqrt{\frac{68.4814}{8 - 2}}$ .

The coefficient of determination is 0.76, found by  $(-0.874)^2$ . Seventy-six percent of the variation in crimes can be explained by the variation in police.

27. The standard error of estimate is 0.913, found by  $\sqrt{\frac{6.667}{10 - 2}}$ .

The coefficient of determination is 0.82, found by  $29.733/36.4$ . Eighty-two percent of the variation in kilowatt hours can be explained by the variation in the number of rooms.

29. a.  $r^2 = \frac{1,000}{1,500} = .6667$

b.  $r = \sqrt{.6667} = .8165$

c.  $s_{y-x} = \sqrt{\frac{500}{13}} = 6.2017$

31. a.  $6.308 \pm (3.182)(.993)\sqrt{.2 + \frac{(7 - 5.6)^2}{29.2}}$

$$= 6.308 \pm 1.633$$

$$= [4.675, 7.941]$$

b.  $6.308 \pm (3.182)(.993)\sqrt{1 + 1/5 + .0671}$

$$= [2.751, 9.865]$$

33. a. 4.2939, 6.3721

b. 2.9854, 7.6806

35. The correlation between the two variables is 0.298.

By squaring  $x$ , the correlation increases to .998.

37.  $H_0: \rho \leq 0$ ;  $H_1: \rho > 0$ . Reject  $H_0$  if  $t > 1.714$ .

$$t = \frac{.94\sqrt{25 - 2}}{\sqrt{1 - (.94)^2}} = 13.213$$

Reject  $H_0$ . There is a positive correlation between passengers and weight of luggage.

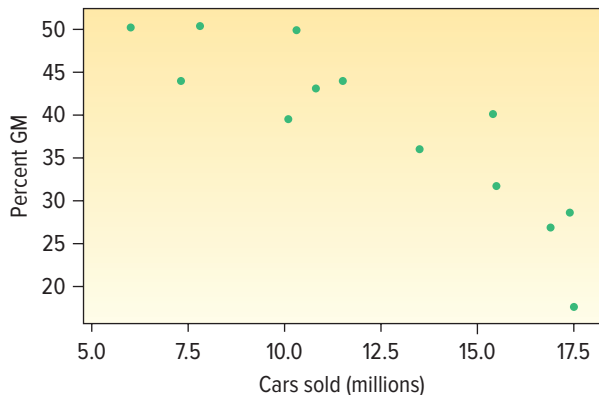
39.  $H_0: \rho \leq 0$ ;  $H_1: \rho > 0$ . Reject  $H_0$  if  $t > 2.764$ .

$$t = \frac{.47\sqrt{12 - 2}}{\sqrt{1 - (.47)^2}} = 1.684$$

Do not reject  $H_0$ . There is not a positive correlation between engine size and performance.  $p$ -value is greater than .05 but less than .10.

41. a. The sales volume is inversely related to their market share.

Scatter Plot of Percent GM vs. Cars Sold (millions)



- b. The correlation coefficient is  $-0.876$ ; there is an inverse linear relationship between the two variables.

- c.  $H_0: \rho \geq 0$   $H_1: \rho < 0$  Reject  $H_0$  if  $t < -2.681$ .  $df = 12$

$$t = \frac{-0.876 \sqrt{14 - 2}}{\sqrt{1 - (-0.876)^2}} = -6.29$$

Reject  $H_0$ . There is a negative correlation between cars sold and market share.

- d. 76.7%, found by  $(-0.876)^2$ , of the variation in market share is accounted for by variation in cars sold.

43. a.  $r = -0.024$

- b. The coefficient of determination is 0.00058, found by squaring  $(-0.024)$ .

- c.  $H_0: \rho \geq 0$   $H_1: \rho < 0$  Reject  $H_0$  if  $t < -1.697$

$$t = \frac{-0.024 \sqrt{32 - 2}}{\sqrt{1 - (-0.024)^2}} = -0.13$$

Reject  $H_0$ . There is a negative correlation between points scored and points allowed.

- d. For the National conference (NFC):  $H_0: \rho \geq 0$   $H_1: \rho < 0$   
Reject  $H_0$  if  $t < -1.761$ .

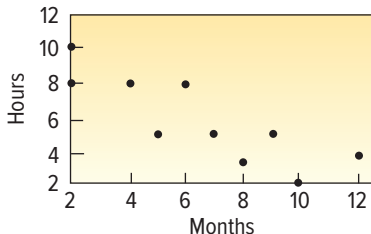
$$t = \frac{-0.139 \sqrt{16 - 2}}{\sqrt{1 - (-0.139)^2}} = -0.53$$

Do not reject  $H_0$ . We cannot say there is a negative correlation between points scored and points allowed in the NFC. For the American conference (AFC):  $H_0: \rho \geq 0$   $H_1: \rho < 0$   
Reject  $H_0$  if  $t < -1.761$ .

$$t = \frac{-0.292 \sqrt{16 - 2}}{\sqrt{1 - (-0.292)^2}} = -1.142$$

Do not reject  $H_0$ . We cannot say there is a negative correlation between points scored and points allowed in the AFC.

45. a.



There is an inverse relationship between the variables. As the months owned increase, the number of hours exercised decreases.

- b.  $r = -0.827$

- c.  $H_0: \rho \geq 0$ ;  $H_1: \rho < 0$ . Reject  $H_0$  if  $t < -2.896$ .

$$t = \frac{-0.827 \sqrt{10 - 2}}{\sqrt{1 - (-0.827)^2}} = -4.16$$

Reject  $H_0$ . There is a negative association between months owned and hours exercised.

47. a. Median age and population are directly related.

$$b. r = \frac{11.93418}{(10 - 1)(2.207)(1.330)} = 0.452$$

- c. The slope of 0.272 indicates that for each increase of 1 million in the population, the median age increases on average by 0.272 year.

- d. The median age is 32.08 years, found by  $31.4 + 0.272(2.5)$ .

- e. The  $p$ -value (0.190) for the population variable is greater than, say, .05. A test for significance of that coefficient would fail to be rejected. In other words, it is possible the population coefficient is zero.

- f.  $H_0: \rho = 0$   $H_1: \rho \neq 0$  Reject  $H_0$  if  $t$  is not between  $-1.86$  and  $1.86$ .

$$df = 8 \quad t = \frac{0.452 \sqrt{10 - 2}}{\sqrt{1 - (0.452)^2}} = 1.433 \text{ Do not reject } H_0.$$

There may be no relationship between age and population.

49. a.  $b = -0.4667$ ,  $a = 11.2358$

- b.  $\hat{y} = 11.2358 - 0.4667(7.0) = 7.9689$

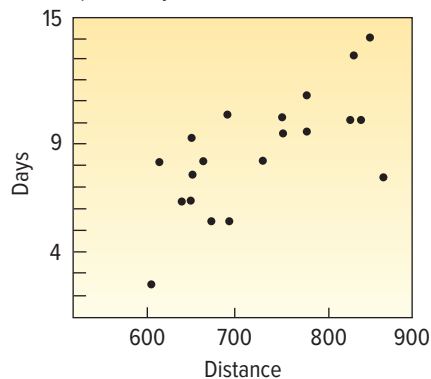
$$c. 7.9689 \pm (2.160)(1.114) \sqrt{1 + \frac{1}{15} + \frac{(7 - 7.1333)^2}{73.7333}}$$

$$= 7.9689 \pm 2.4854$$

$$= [5.4835, 10.4543]$$

- d.  $R^2 = 0.499$ . Nearly 50% of the variation in the amount of the bid is explained by the number of bidders.

51. a.



There appears to be a relationship between the two variables. As the distance increases, so does the shipping time.

- b.  $r = 0.692$

$H_0: \rho \leq 0$ ;  $H_1: \rho > 0$ . Reject  $H_0$  if  $t > 1.734$ .

$$t = \frac{0.692 \sqrt{20 - 2}}{\sqrt{1 - (0.692)^2}} = 4.067$$

$H_0$  is rejected. There is a positive association between shipping distance and shipping time.

- c.  $R^2 = 0.479$ . Nearly half of the variation in shipping time is explained by shipping distance.

- d.  $s_{y,x} = 2.004$

- e. No, the  $R^2$  is not large enough to support an accurate prediction of shipping time.

53. a.  $b = 2.41$

$a = 26.8$

The regression equation is: Price =  $26.8 + 2.41 \times$  Dividend. For each additional dollar of dividend, the price increases by \$2.41.

- b. To test the significance of the slope, we use  $n - 2$ , or  $30 - 2 = 28$  degrees of freedom. For a 0.05 level of significance, the critical values are  $-2.048$  and  $2.048$ . The  $t$ -test statistic is  $t = \frac{b - 0}{s_b} = \frac{2.408}{0.328} = 7.34$ . We reject the null hypothesis that the slope is equal to zero.

c.  $R^2 = \frac{5,057.6}{7,682.7} = 0.658$  Thus, 65.8% of the variation in price is explained by the dividend.

- d.  $r = \sqrt{0.658} = 0.811$   $H_0: \rho \leq 0$   $H_1: \rho > 0$   
At the 5% level, reject  $H_0$  when  $t > 1.701$ .

$$t = \frac{0.811\sqrt{30 - 2}}{\sqrt{1 - (0.811)^2}} = 7.34$$

Thus,  $H_0$  is rejected. The population correlation is positive.

55. a. 35  
b.  $s_{y-x} = \sqrt{29,778,406} = 5,456.96$   
c.  $r^2 = \frac{13,548,662,082}{14,531,349,474} = 0.932$   
d.  $r = \sqrt{0.932} = 0.966$   
e.  $H_0: \rho \leq 0$ ,  $H_1: \rho > 0$ . Reject  $H_0$  if  $t > 1.692$ .

$$t = \frac{.966\sqrt{35 - 2}}{\sqrt{1 - (.966)^2}} = 21.46$$

Reject  $H_0$ . There is a direct relationship between size of the house and its market value.

57. a. The regression equation is  $\text{Price} = -386.5 + 704.0 \text{ Speed}$ .  
b. The computers 2, 3, and 10 have errors in excess of \$200.00.  
c. The correlation of Speed and Price is 0.835.  
 $H_0: \rho \leq 0$   $H_1: \rho > 0$  Reject  $H_0$  if  $t > 1.8125$ .

$$t = \frac{0.835\sqrt{12 - 2}}{\sqrt{1 - (0.835)^2}} = 4.799$$

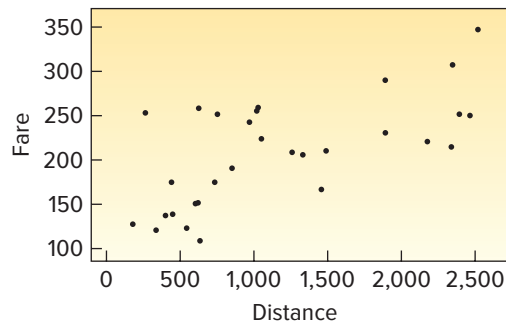
Reject  $H_0$ . It is reasonable to say the population correlation is positive.

59. a.  $r = .987$ ,  $H_0: \rho \leq 0$ ,  $H_1: \rho > 0$ . Reject  $H_0$  if  $t > 1.746$ .

$$t = \frac{.987\sqrt{18 - 2}}{\sqrt{1 - (.987)^2}} = 24.564$$

- b.  $\hat{y} = -29.7 + 22.93x$ ; an additional cup increases the dog's weights by almost 23 pounds.  
c. Dog number 4 is an overeater.

61. a. Scatter Diagram of Fares and Distances



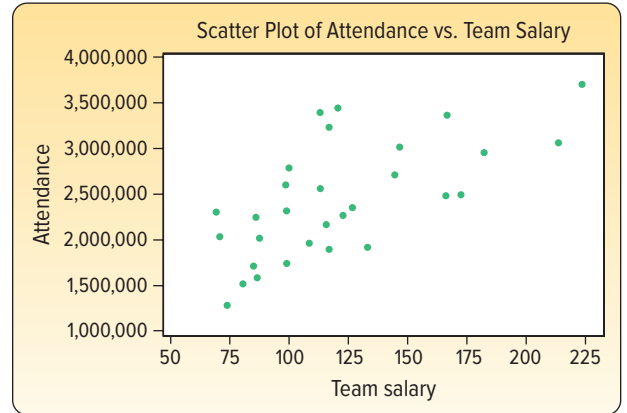
The relationship is direct. Fares increase for longer flights.

- b. The correlation of Distance and Fare is 0.656.  
 $H_0: \rho \leq 0$   $H_1: \rho > 0$  Reject  $H_0$  if  $t > 1.701$ .  $df = 28$

$$t = \frac{0.656\sqrt{30 - 2}}{\sqrt{1 - (0.656)^2}} = 4.599$$

Reject  $H_0$ . There is a significant positive correlation between fares and distances.

- c. 43%, found by  $(0.656)^2$ , of the variation in fares is explained by the variation in distance.  
d. The regression equation is  $\text{Fare} = 147.08 + 0.05265 \text{ Distance}$ . Each additional mile adds \$0.05265 to the fare. A 1,500-mile flight would cost \$226.06, found by  $\$147.08 + 0.05265(1,500)$ .  
e. A flight of 4,218 miles is outside the range of the sampled data, so the regression equation may not be useful.  
63. a. There does seem to be a direct relationship between the variables.



- b. Expected Attendance with a salary of \$100 million is 2,224,268, found by  $1,208,968 + 10,153(100)$ .  
c. Increasing the salary by \$30 million will increase attendance by 304,590 on average, found by  $10,153(30)$ . This is also the difference between the expected attendance with a salary of \$130 million and the expected attendance of 100 million.  
d. The regression output from Excel is below.

| SUMMARY OUTPUT        |              |                    |                   |         |                |
|-----------------------|--------------|--------------------|-------------------|---------|----------------|
| Regression Statistics |              |                    |                   |         |                |
| Multiple R            |              | 0.652              |                   |         |                |
| R Square              |              | 0.425              |                   |         |                |
| Adjusted R Square     |              | 0.404              |                   |         |                |
| Standard Error        |              | 484846.605         |                   |         |                |
| Observations          |              | 30.000             |                   |         |                |
| ANOVA                 |              |                    |                   |         |                |
|                       | df           | SS                 | MS                | F       | Significance F |
| Regression            | 1            | 4863338397693.920  | 4863338397693.920 | 20.688  | 0.000          |
| Residual              | 28           | 6582134442815.280  | 235076230100.546  |         |                |
| Total                 | 29           | 11445472840509.200 |                   |         |                |
|                       | Coefficients | Standard Error     | t Stat            | P-value |                |
| Intercept             | 1208968.430  | 284472.197         | 4.250             | 0.000   |                |
| Team Salary           | 10152.694    | 2232.124           | 4.548             | 0.000   |                |

$H_0: \beta \leq 0$   $H_1: \beta > 0$   $df = n - 2 = 30 - 2 = 28$   
Reject  $H_0$  if  $t > 1.701$ .  $t = 10152.694/2232.124 = 4.548$ .  
Reject  $H_0$  and conclude the slope is positive.

e. 0.4249, or 42.49%, of the variation in attendance is explained by variation in salary.

f. The correlation between attendance and batting average is 0.1472.

$H_0: \rho \leq 0$     $H_1: \rho > 0$    At the 5% level, reject  $H_0$  if  $t > 1.701$ .

$$t = \frac{0.1472\sqrt{30-2}}{\sqrt{1-(0.1472)^2}} = 0.787$$

Fail to reject  $H_0$ .

$p$ -value = 0.4377

The batting average and attendance are not positively correlated.

The correlation between attendance and ERA is  $-0.4745$ .

The correlation between attendance and ERA is stronger than the correlation between attendance and batting average.

$H_0: \rho \geq 0$     $H_1: \rho < 0$    At the 5% level, reject  $H_0$  if  $t < -1.701$ .

$$t = \frac{-0.4745\sqrt{30-2}}{\sqrt{1-(-0.4745)^2}} = -2.8524$$

Fail to reject  $H_0$ .

$p$ -value = 0.0081

The ERA and attendance are not negatively correlated.

## CHAPTER 14

1. a. Multiple regression equation  
b. The  $y$ -intercept  
c.  $\hat{y} = 64,100 + 0.394(796,000) + 9.6(6,940) - 11,600(6.0) = \$374,748$
3. a. 497.736, found by  
 $\hat{y} = 16.24 + 0.017(18) + 0.0028(26,500) + 42(3) + 0.0012(156,000) + 0.19(14) + 26.8(2.5)$   
b. Two more social activities. Income added only 28 to the index; social activities added 53.6.
5. a.  $s_{y-12} = \sqrt{\frac{SSE}{n-(k+1)}} = \sqrt{\frac{583.693}{65-(2+1)}} = \sqrt{9.414} = 3.068$   
95% of the residuals will be between  $\pm 6.136$ , found by  $2(3.068)$ .  
b.  $R^2 = \frac{SSR}{SS \text{ total}} = \frac{77.907}{661.6} = .118$   
The independent variables explain 11.8% of the variation.  
c.  $R_{adj}^2 = 1 - \frac{\frac{SSE}{n-(k+1)}}{\frac{SS \text{ total}}{n-1}} = 1 - \frac{\frac{583.693}{65-(2+1)}}{\frac{661.6}{65-1}} = 1 - \frac{9.414}{10.3375} = 1 - .911 = .089$
7. a.  $\hat{y} = 84.998 + 2.391x_1 - 0.4086x_2$   
b. 90.0674, found by  $\hat{y} = 84.998 + 2.391(4) - 0.4086(11)$   
c.  $n = 65$  and  $k = 2$   
d.  $H_0: \beta_1 = \beta_2 = 0$     $H_1$ : Not all  $\beta$ s are 0  
Reject  $H_0$  if  $F > 3.15$ .  
 $F = 4.14$ , reject  $H_0$ . Not all net regression coefficients equal zero.  
e. For  $x_1$    For  $x_2$   
 $H_0: \beta_1 = 0$     $H_0: \beta_2 = 0$   
 $H_1: \beta_1 \neq 0$     $H_1: \beta_2 \neq 0$   
 $t = 1.99$     $t = -2.38$   
Reject  $H_0$  if  $t > 2.0$  or  $t < -2.0$ .  
Delete variable 1 and keep 2.  
f. The regression analysis should be repeated with only  $x_2$  as the independent variable.

9. a. The regression equation is: Performance = 29.3 + 5.22 Aptitude + 22.1 Union

| Predictor | Coef   | SE Coef | T    | P     |
|-----------|--------|---------|------|-------|
| Constant  | 29.28  | 12.77   | 2.29 | 0.041 |
| Aptitude  | 5.222  | 1.702   | 3.07 | 0.010 |
| Union     | 22.135 | 8.852   | 2.50 | 0.028 |

$S = 16.9166$     $R\text{-Sq} = 53.3\%$     $R\text{-Sq (adj)} = 45.5\%$

Analysis of Variance

| Source         | DF | SS     | MS     | F    | P     |
|----------------|----|--------|--------|------|-------|
| Regression     | 2  | 3919.3 | 1959.6 | 6.85 | 0.010 |
| Residual Error | 12 | 3434.0 | 286.2  |      |       |
| Total          | 14 | 7353.3 |        |      |       |

- b. These variables are effective in predicting performance. They explain 53.3% of the variation in performance. In particular, union membership increases the typical performance by 22.1.
- c.  $H_0: \beta_2 = 0$     $H_1: \beta_2 \neq 0$   
Reject  $H_0$  if  $t < -2.179$  or  $t > 2.179$ . Since 2.50 is greater than 2.179, we reject the null hypothesis and conclude that union membership is significant and should be included.
11. a.  $n = 40$   
b. 4  
c.  $R^2 = \frac{750}{1,250} = .60$   
d.  $s_{y-1234} = \sqrt{500/35} = 3.7796$   
e.  $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$   
 $H_1$ : Not all the  $\beta$ s equal zero.  
 $H_0$  is rejected if  $F > 2.65$ .  
 $F = \frac{750/4}{500/35} = 13.125$   
 $H_0$  is rejected. At least one  $\beta_i$  does not equal zero.
13. a.  $n = 26$   
b.  $R^2 = 100/140 = .7143$   
c. 1.4142, found by  $\sqrt{2}$   
d.  $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$   
 $H_1$ : Not all the  $\beta$ s are 0.  
 $H_0$  is rejected if  $F > 2.71$ .  
Computed  $F = 10.0$ . Reject  $H_0$ . At least one regression coefficient is not zero.  
e.  $H_0$  is rejected in each case if  $t < -2.086$  or  $t > 2.086$ .  
 $x_1$  and  $x_5$  should be dropped.
15. a. \$28,000  
b.  $R^2 = \frac{SSR}{SS \text{ total}} = \frac{3,050}{5,250} = .5809$   
c. 9.199, found by  $\sqrt{84.62}$   
d.  $H_0$  is rejected if  $F > 2.97$  (approximately).  
Computed  $F = \frac{1,016.67}{84.62} = 12.01$   
 $H_0$  is rejected. At least one regression coefficient is not zero.  
e. If computed  $t$  is to the left of  $-2.056$  or to the right of 2.056, the null hypothesis in each of these cases is rejected. Computed  $t$  for  $x_2$  and  $x_3$  exceed the critical value. Thus, "population" and "advertising expenses" should be retained and "number of competitors,"  $x_1$ , dropped.
17. a. The strongest correlation is between High School GPA and Paralegal GPA. No problem with multicollinearity.  
b.  $R^2 = \frac{4.3595}{5.0631} = .8610$   
c.  $H_0$  is rejected if  $F > 5.41$ .  
Computed  $F = \frac{1.4532}{0.1407} = 10.328$   
At least one coefficient is not zero.  
d. Any  $H_0$  is rejected if  $t < -2.571$  or  $t > 2.571$ . It appears that only High School GPA is significant. Verbal and math could be eliminated.

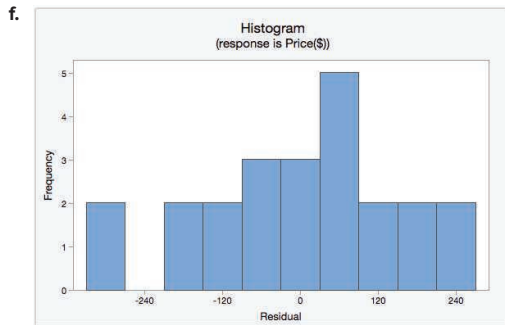
e.  $R^2 = \frac{4.2061}{5.0631} = .8307$

$R^2$  has only been reduced .0303.

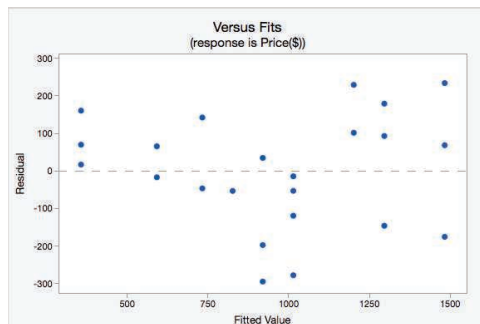
- f. The residuals appear slightly skewed (positive) but acceptable.  
 g. There does not seem to be a problem with the plot.

19. a. The correlation of Screen and Price is 0.893. So there does appear to be a linear relationship between the two.  
 b. Price is the "dependent" variable.  
 c. The regression equation is  $\text{Price} = -1242.1 + 50.671(\text{screen size})$ . For each inch increase in screen size, the price increases \$50.671 on average.  
 d. Using indicator or dummy variables for Sony and Sharp, note that when the values for Sony and Sharp are zero, then the television is Samsung. When this is true, the regression equation is  $\text{Price} = -1154.103 + 47.06(\text{screen size})$ . It is significant, i.e.,  $p$ -value is less than 0.05. When Sony = 1 and Sharp is equal to 0, the Sony coefficient is 190.72 and is significant. It shows that, on average, a Sony TV is \$190.72 more expensive than a Samsung. When Sharp = 1 and Sony is equal to 0, the Sharp coefficient is 7.55. However, it is not significantly different from zero with a  $p$ -value of 0.93. So, Sharp and Samsung TVs are about the same price.  
 e.

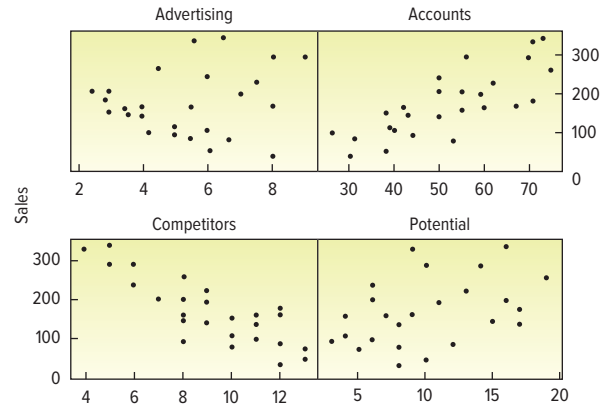
|           | Coefficients | Standard Error | t Stat    | P-value  |
|-----------|--------------|----------------|-----------|----------|
| Intercept | -1154.103299 | 245.9847351    | -4.691768 | 0.000159 |
| Sharp     | 7.549207239  | 85.8232053     | 0.087962  | 0.930827 |
| Sony      | 190.7193306  | 84.40199954    | 2.259654  | 0.035784 |
| Screen    | 47.05997758  | 5.415186736    | 8.69037   | 4.8E-08  |



- f. There is no apparent relationship in the residuals, but the residual variation may be increasing with larger fitted values.



21. a. Scatter Diagram of Sales vs. Advertising, Accounts, Competitors, Potential



Sales seem to fall with the number of competitors and rise with the number of accounts and potential.

- b. Pearson correlations

|             | Sales  | Advertising | Accounts | Competitors |
|-------------|--------|-------------|----------|-------------|
| Advertising | 0.159  |             |          |             |
| Accounts    | 0.783  | 0.173       |          |             |
| Competitors | -0.833 | -0.038      | -0.324   |             |
| Potential   | 0.407  | -0.071      | 0.468    | -0.202      |

The number of accounts and the market potential are moderately correlated.

- c. The regression equation is:

$$\text{Sales} = 178 + 1.81 \text{ Advertising} + 3.32 \text{ Accounts} - 21.2 \text{ Competitors} + 0.325 \text{ Potential}$$

| Predictor   | Coef     | SE Coef | T      | P     |
|-------------|----------|---------|--------|-------|
| Constant    | 178.32   | 12.96   | 13.76  | 0.000 |
| Advertising | 1.807    | 1.081   | 1.67   | 0.109 |
| Accounts    | 3.3178   | 0.1629  | 20.37  | 0.000 |
| Competitors | -21.1850 | 0.7879  | -26.89 | 0.000 |
| Potential   | 0.3245   | 0.4678  | 0.69   | 0.495 |

$$S = 9.60441 \quad R\text{-Sq} = 98.9\% \quad R\text{-Sq(adj)} = 98.7\%$$

Analysis of Variance

| Source         | DF | SS     | MS    | F      | P     |
|----------------|----|--------|-------|--------|-------|
| Regression     | 4  | 176777 | 44194 | 479.10 | 0.000 |
| Residual Error | 21 | 1937   | 92    |        |       |
| Total          | 25 | 178714 |       |        |       |

The computed  $F$ -value is quite large, so we can reject the null hypothesis that all of the regression coefficients are zero. We conclude that some of the independent variables are effective in explaining sales.

- d. Market potential and advertising have large  $p$ -values (0.495 and 0.109, respectively). You would probably drop them.

- e. If you omit potential, the regression equation is:

$$\text{Sales} = 180 + 1.68 \text{ Advertising} + 3.37 \text{ Accounts} - 21.2 \text{ Competitors}$$

| Predictor   | Coef     | SE Coef | T      | P     |
|-------------|----------|---------|--------|-------|
| Constant    | 179.84   | 12.62   | 14.25  | 0.000 |
| Advertising | 1.677    | 1.052   | 1.59   | 0.125 |
| Accounts    | 3.3694   | 0.1432  | 23.52  | 0.000 |
| Competitors | -21.2165 | 0.7773  | -27.30 | 0.000 |

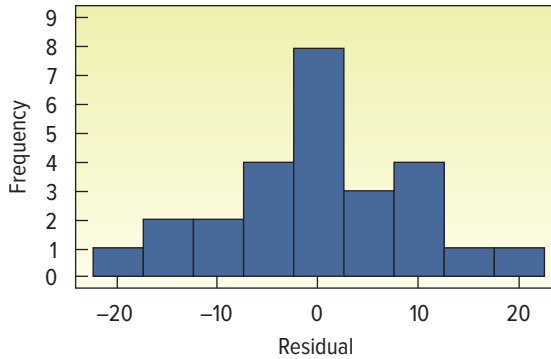
Now advertising is not significant. That would also lead you to cut out the advertising variable and report that the polished regression equation is:  $\text{Sales} = 187 + 3.41 \text{ Accounts} - 21.2 \text{ Competitors}$

| Predictor   | Coef     | SE Coef | T      | P     |
|-------------|----------|---------|--------|-------|
| Constant    | 186.69   | 12.26   | 15.23  | 0.000 |
| Accounts    | 3.4081   | 0.1458  | 23.37  | 0.000 |
| Competitors | -21.1930 | 0.8028  | -26.40 | 0.000 |



f.

Histogram of the Residuals  
(response is Sales)



The histogram looks to be normal. There are no problems shown in this plot.

- g. The variance inflation factor for both variables is 11. They are less than 10. There are no troubles, as this value indicates the independent variables are not strongly correlated with each other.

23. The computer output is:

| Predictor | Coef   | StDev | t-ratio | p     |
|-----------|--------|-------|---------|-------|
| Constant  | 651.9  | 345.3 | 1.89    | 0.071 |
| Service   | 13.422 | 5.125 | 2.62    | 0.015 |
| Age       | -6.710 | 6.349 | -1.06   | 0.301 |
| Gender    | 205.65 | 90.27 | 2.28    | 0.032 |
| Job       | -33.45 | 89.55 | -0.37   | 0.712 |

| Analysis of Variance |    |         |        |      |       |
|----------------------|----|---------|--------|------|-------|
| SOURCE               | DF | SS      | MS     | F    | p     |
| Regression           | 4  | 1066830 | 266708 | 4.77 | 0.005 |
| Error                | 25 | 1398651 | 55946  |      |       |
| Total                | 29 | 2465481 |        |      |       |

- a.  $\hat{y} = 651.9 + 13.422x_1 - 6.710x_2 + 205.65x_3 - 33.45x_4$   
 b.  $R^2 = .433$ , which is somewhat low for this type of study.  
 c.  $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ ;  $H_1$ : Not all  $\beta$ s equal zero.  
 Reject  $H_0$  if  $F > 2.76$ .

$$F = \frac{1,066,830/4}{1,398,651/25} = 4.77$$

$H_0$  is rejected. Not all the  $\beta$ s equal 0.

- d. Using the .05 significance level, reject the hypothesis that the regression coefficient is 0 if  $t < -2.060$  or  $t > 2.060$ . Service and gender should remain in the analyses; age and job should be dropped.  
 e. Following is the computer output using the independent variables service and gender.

| Predictor | Coef   | StDev | t-ratio | p     |
|-----------|--------|-------|---------|-------|
| Constant  | 784.2  | 316.8 | 2.48    | 0.020 |
| Service   | 9.021  | 3.106 | 2.90    | 0.007 |
| Gender    | 224.41 | 87.35 | 2.57    | 0.016 |

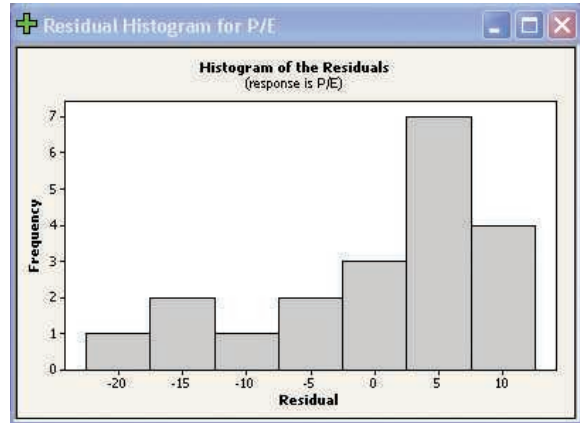
  

| Analysis of Variance |    |         |        |      |       |
|----------------------|----|---------|--------|------|-------|
| SOURCE               | DF | SS      | MS     | F    | p     |
| Regression           | 2  | 998779  | 499389 | 9.19 | 0.001 |
| Error                | 27 | 1466703 | 54322  |      |       |
| Total                | 29 | 2465481 |        |      |       |

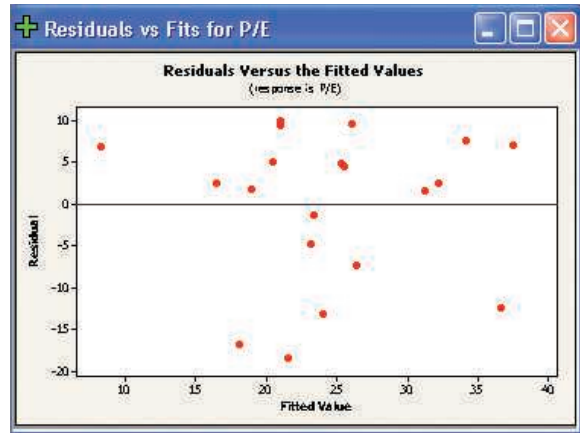
A man earns \$224 more per month than a woman. The difference between management and engineering positions is not significant.

25. a.  $\hat{y} = 29.913 - 5.324x_1 + 1.449x_2$   
 b. EPS is ( $t = -3.26$ ,  $p$ -value = .005). Yield is not ( $t = 0.81$ ,  $p$ -value = .431).

- c. An increase of 1 in EPS results in a decline of 5.324 in P/E.  
 d. Stock number 2 is undervalued.  
 e. Below is a residual plot. It does *not* appear to follow the normal distribution.



- f. There does not seem to be a problem with the plot of the residuals versus the fitted values.



- g. The correlation between yield and EPS is not a problem. No problem with multicollinearity.

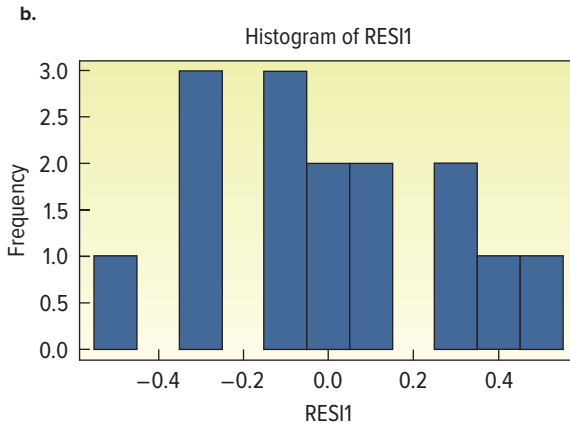
|       | P/E    | EPS  |
|-------|--------|------|
| EPS   | -0.602 |      |
| Yield | .054   | .162 |

27. a. The regression equation is  
 Sales (000) = 1.02 + 0.0829 Infomercials

| Predictor    | Coef    | SE Coef | T    | P     |
|--------------|---------|---------|------|-------|
| Constant     | 1.0188  | 0.3105  | 3.28 | 0.006 |
| Infomercials | 0.08291 | 0.01680 | 4.94 | 0.000 |

| Analysis of Variance |    |        |        |       |       |
|----------------------|----|--------|--------|-------|-------|
| Source               | DF | SS     | MS     | F     | P     |
| Regression           | 1  | 2.3214 | 2.3214 | 24.36 | 0.000 |
| Residual Error       | 13 | 1.2386 | 0.0953 |       |       |
| Total                | 14 | 3.5600 |        |       |       |

The global test demonstrates there is a relationship between sales and the number of infomercials.



The residuals appear to follow the normal distribution.

29. a. The correlation matrix is as follows:

|                    | Price | Bedrooms | Size (square feet) | Baths  | Days on Market |
|--------------------|-------|----------|--------------------|--------|----------------|
| Price              | 1.000 |          |                    |        |                |
| Bedrooms           | 0.844 | 1.000    |                    |        |                |
| Size (square feet) | 0.952 | 0.877    | 1.000              |        |                |
| Baths              | 0.825 | 0.985    | 0.851              | 1.000  |                |
| Days on Market     | 0.185 | 0.002    | 0.159              | -0.002 | 1              |

There are strong, positive correlations between “Price” and the independent variables “Bedrooms,” “Size,” and “Baths.” There appears to be no relationship between “Price” and “Days on Market.” The correlations among the independent variables are very strong. So, there would be a high degree of multicollinearity in a multiple regression equation if all the variables were included. We will need to be careful in selecting the best independent variable to predict price.

b.

| SUMMARY OUTPUT        |              |                |           |           |                |
|-----------------------|--------------|----------------|-----------|-----------|----------------|
| Regression Statistics |              |                |           |           |                |
| Multiple R            | 0.952        |                |           |           |                |
| R Square              | 0.905        |                |           |           |                |
| Adjusted R Square     | 0.905        |                |           |           |                |
| Standard Error        | 49655.822    |                |           |           |                |
| Observations          | 105.000      |                |           |           |                |
| ANOVA                 |              |                |           |           |                |
|                       | df           | SS             | MS        | F         | Significance F |
| Regression            | 1            | 2.432E+12      | 2.432E+12 | 9.862E+02 | 1.46136E-54    |
| Residual              | 103          | 2.540E+11      | 2.466E+09 |           |                |
| Total                 | 104          | 2.686E+12      |           |           |                |
|                       | Coefficients | Standard Error | t Stat    | P-value   |                |
| Intercept             | -15775.995   | 12821.967      | -1.230    | 0.221     |                |
| Size (square feet)    | 108.364      | 3.451          | 31.405    | 0.000     |                |

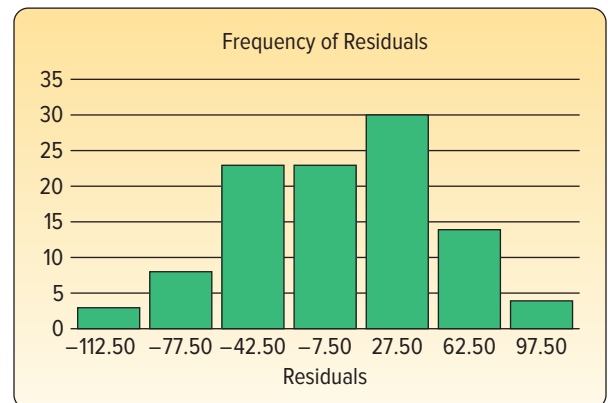
The regression analysis shows a significant relationship between price and house size. The  $p$ -value of the  $F$ -statistic is 0.00, so the null hypothesis of “no relationship” is rejected. Also, the  $p$ -value associated with the regression coefficient of “size” is 0.000. Therefore, this coefficient is clearly different from zero. The regression equation is: Price =  $-15775.995 + 108.364$  Size.

In terms of pricing, the regression equation suggests that houses are priced at about \$108 per square foot.

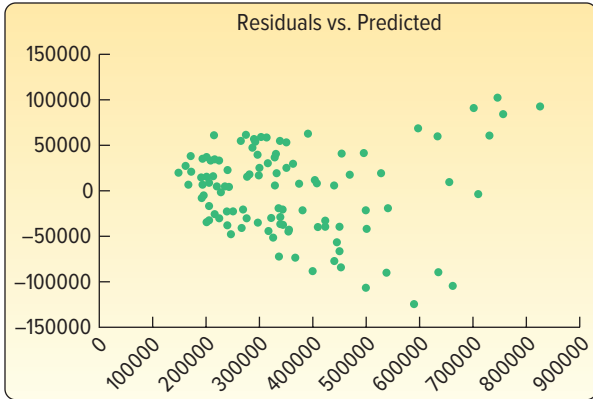
c. The regression analyses of price and size with the qualitative variables pool and garage follow. The results show that the variable “pool” is statistically significant in the equation. The regression coefficient indicates that if a house has a pool, it adds about \$28,575 to the price. The analysis of including “garage” to the analysis indicates that it does not affect the pricing of the house. Adding pool to the regression equation increases the  $R$ -square by about 1%.

| SUMMARY OUTPUT        |        |                  |                  |        |                |
|-----------------------|--------|------------------|------------------|--------|----------------|
| Regression Statistics |        |                  |                  |        |                |
| Multiple R            |        | 0.955            |                  |        |                |
| R Square              |        | 0.913            |                  |        |                |
| Adjusted R Square     |        | 0.911            |                  |        |                |
| Standard Error        |        | 47914.856        |                  |        |                |
| Observations          |        | 105              |                  |        |                |
| ANOVA                 |        |                  |                  |        |                |
|                       | df     | SS               | MS               | F      | Significance F |
| Regression            | 2.00   | 2451577033207.43 | 1225788516603.72 | 533.92 | 0.00           |
| Residual              | 102.00 | 234175013207.24  | 2295833462.82    |        |                |
| Total                 | 104.00 | 2685752046414.68 |                  |        |                |
|                       |        | Coefficients     | Standard Error   | t Stat | P-value        |
| Intercept             |        | -34640.573       | 13941.203        | -2.485 | 0.015          |
| Size (square feet)    |        | 108.547          | 3.330            | 32.595 | 0.000          |
| Pool (yes is 1)       |        | 28575.145        | 9732.223         | 2.936  | 0.004          |

d. The following histogram was developed using the residuals from part (c). The normality assumption is reasonable.



e. The following scatter diagram is based on the residuals in part (c) with the predicted dependent variable on the horizontal axis and residuals on the vertical axis. There does appear that the variance of the residuals increases with higher values of the predicted price. You can experiment with transformations such as the log of price or the square root of price and observe the changes in the graphs of residuals. Note that the transformations will make the interpretation of the regression equation more difficult.



Last, it is possible that maintenance costs are different for diesel versus gasoline engines. So, adding this variable to the analysis shows:

| SUMMARY OUTPUT               |           |                     |                       |               |                       |
|------------------------------|-----------|---------------------|-----------------------|---------------|-----------------------|
| <b>Regression Statistics</b> |           |                     |                       |               |                       |
| Multiple R                   |           | 0.960               |                       |               |                       |
| R Square                     |           | 0.922               |                       |               |                       |
| Adjusted R Square            |           | 0.920               |                       |               |                       |
| Standard Error               |           | 658.369             |                       |               |                       |
| Observations                 |           | 80                  |                       |               |                       |
| <b>ANOVA</b>                 |           |                     |                       |               |                       |
|                              | <i>df</i> | <i>SS</i>           | <i>MS</i>             | <i>F</i>      | <i>Significance F</i> |
| Regression                   | 2         | 396072093.763       | 198036046.881         | 456.884       | 0.000                 |
| Residual                     | 77        | 33375590.225        | 433449.224            |               |                       |
| Total                        | 79        | 429447683.988       |                       |               |                       |
|                              |           | <b>Coefficients</b> | <b>Standard Error</b> | <b>t Stat</b> | <b>P-value</b>        |
| Intercept                    |           | -1028.539           | 213.761               | -4.812        | 0.000                 |
| Age (years)                  |           | 644.528             | 27.157                | 23.733        | 0.000                 |
| Engine Type (0=diesel)       |           | 3190.481            | 156.100               | 20.439        | 0.000                 |

31. a.

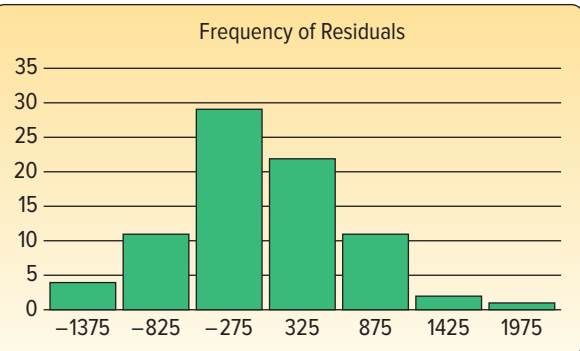
|                         | Maintenance Cost (\$) | Age (years)  | Odometer Miles | Miles since Last Maintenance |
|-------------------------|-----------------------|--------------|----------------|------------------------------|
| Maintenance Cost (\$)   | 1                     |              |                |                              |
| Age (years)             | 0.710194278           | 1            |                |                              |
| Odometer Miles          | 0.700439797           | 0.990675674  | 1              |                              |
| Miles since Last Maint. | -0.160275988          | -0.140196856 | -0.118982823   | 1                            |

The correlation analysis shows that age and odometer miles are positively correlated with cost and that "miles since last maintenance" shows that costs increase with fewer miles between maintenance. The analysis also shows a strong correlation between age and odometer miles. This indicates the strong possibility of multicollinearity if age and odometer miles are included in a regression equation.

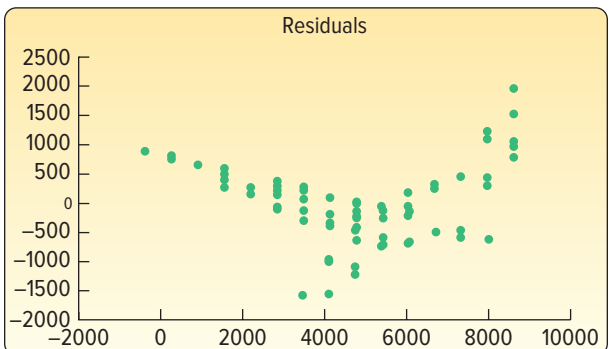
- b. There are a number of analyses to do. When you use age or odometer miles as an independent variable and you review these analyses, both result in significant relationships. However, age has a slightly higher  $R^2$ , which makes it a better choice as the first independent variable. The interpretation of the coefficient using age is bit more useful. That is, we can expect about an average of \$600 increase in maintenance costs for each additional year a bus ages. The results are:

| SUMMARY OUTPUT               |           |                     |                       |               |                       |
|------------------------------|-----------|---------------------|-----------------------|---------------|-----------------------|
| <b>Regression Statistics</b> |           |                     |                       |               |                       |
| Multiple R                   |           | 0.708               |                       |               |                       |
| R Square                     |           | 0.501               |                       |               |                       |
| Adjusted R Square            |           | 0.494               |                       |               |                       |
| Standard Error               |           | 1658.097            |                       |               |                       |
| Observations                 |           | 80                  |                       |               |                       |
| <b>ANOVA</b>                 |           |                     |                       |               |                       |
|                              | <i>df</i> | <i>SS</i>           | <i>MS</i>             | <i>F</i>      | <i>Significance F</i> |
| Regression                   | 1         | 215003471.845       | 215003471.845         | 78.203        | 0.000                 |
| Residual                     | 78        | 214444212.142       | 2749284.771           |               |                       |
| Total                        | 79        | 429447683.988       |                       |               |                       |
|                              |           | <b>Coefficients</b> | <b>Standard Error</b> | <b>t Stat</b> | <b>P-value</b>        |
| Intercept                    |           | 337.297             | 511.372               | 0.660         | 0.511                 |
| Age (years)                  |           | 603.161             | 68.206                | 8.843         | 0.000                 |

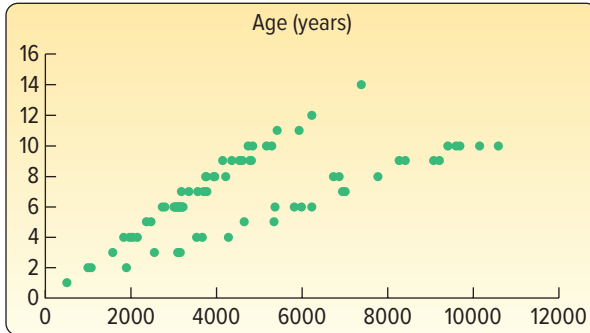
We can also explore including the variable "miles since last maintenance" with age. Your analysis will show that "miles since last maintenance" is not significantly related to costs.



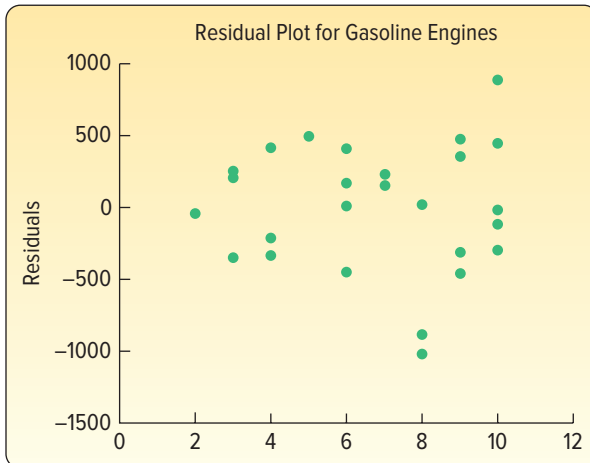
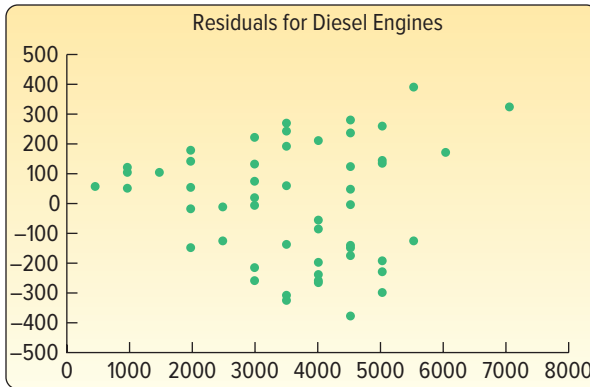
- d. The plot of residuals versus predicted values shows the following. There are clearly patterns in the graph that indicate that the residuals do not follow the assumptions required for the tests of hypotheses.



Let's remember the scatter plot of costs versus age. The graph clearly shows the effect of engine type on costs. So there are essentially two regression equations, depending on the type of engine.



So based on our knowledge of the data, let's create a residual plot of costs for each engine type.



The graphs show a much better distribution of residuals.

## CHAPTER 15

- $H_0$  is rejected if  $z > 1.65$ .
  - 1.09, found by  $z = (0.75 - 0.70) / \sqrt{(0.70 \times 0.30) / 100}$
  - $H_0$  is not rejected.
- Step 1:**  $H_0: \pi = 0.10$     $H_1: \pi \neq 0.10$
  - Step 2:** The 0.01 significance level was chosen.
  - Step 3:** Use the z-statistic, as the binomial distribution can be approximated by the normal distribution as  $n\pi = 30 > 5$  and  $n(1-\pi) = 270 > 5$ .
  - Step 4:** Reject  $H_0$  if  $z > 2.326$ .
  - Step 5:**

$$z = \frac{\{(^{63}/_{300}) - 0.10\}}{\sqrt{\{0.10(0.90)/_{300}\}}} = 6.35$$

Reject  $H_0$ .

**Step 6:** We conclude that the proportion of carpooling cars on the Turnpike is not 10%.

- $H_0: \pi \geq 0.90$     $H_1: \pi < 0.90$
  - $H_0$  is rejected if  $z < -1.28$ .
  - 2.67, found by  $z = (0.82 - 0.90) / \sqrt{(0.90 \times 0.10) / 100}$
  - $H_0$  is rejected. Fewer than 90% of the customers receive their orders in less than 10 minutes.
- $H_0$  is rejected if  $z > 1.65$ .
  - 0.64, found by  $p_c = \frac{70 + 90}{100 + 150}$
  - 1.61, found by

$$z = \frac{0.70 - 0.60}{\sqrt{[(0.64 \times 0.36) / 100] + [(0.64 \times 0.36) / 150]}}$$

d.  $H_0$  is not rejected.

- $H_0: \pi_1 = \pi_2$     $H_1: \pi_1 \neq \pi_2$
  - $H_0$  is rejected if  $z < -1.96$  or  $z > 1.96$ .
  - $p_c = \frac{24 + 40}{400 + 400} = 0.08$
  - 2.09, found by

$$z = \frac{0.06 - 0.10}{\sqrt{[(0.08 \times 0.92) / 400] + [(0.08 \times 0.92) / 400]}}$$

e.  $H_0$  is rejected. The proportion infested is not the same in the two fields.

- $H_0: \pi_d \leq \pi_r$     $H_1: \pi_d > \pi_r$
  - $H_0$  is rejected if  $z > 2.05$ .

$$p_c = \frac{168 + 200}{800 + 1,000} = 0.2044$$

$$z = \frac{0.21 - 0.20}{\sqrt{\frac{(0.2044)(0.7956)}{800} + \frac{(0.2044)(0.7956)}{1,000}}} = 0.52$$

$H_0$  is not rejected. We cannot conclude that a larger proportion of Democrats favor lowering the standards.  $p$ -value = .3015.

- 3
  - 7.815
- Reject  $H_0$  if  $\chi^2 > 5.991$
  - $\chi^2 = \frac{(10 - 20)^2}{20} + \frac{(20 - 20)^2}{20} + \frac{(30 - 20)^2}{20} = 10.0$
  - Reject  $H_0$ . The proportions are not equal.
- $H_0$ : The outcomes are the same;  $H_1$ : The outcomes are not the same. Reject  $H_0$  if  $\chi^2 > 9.236$ .

$$\chi^2 = \frac{(3 - 5)^2}{5} + \dots + \frac{(7 - 5)^2}{5} = 7.60$$

Do not reject  $H_0$ . Cannot reject  $H_0$  when outcomes are the same.

- $H_0$ : There is no difference in the proportions.
  - $H_1$ : There is a difference in the proportions.
  - Reject  $H_0$  if  $\chi^2 > 15.086$ .

$$\chi^2 = \frac{(47 - 40)^2}{40} + \dots + \frac{(34 - 40)^2}{40} = 3.400$$

Do not reject  $H_0$ . There is no difference in the proportions.

21. a. Reject  $H_0$  if  $\chi^2 > 9.210$ .  
 b.  $\chi^2 = \frac{(30 - 24)^2}{24} + \frac{(20 - 24)^2}{24} + \frac{(10 - 12)^2}{12} = 2.50$   
 c. Do not reject  $H_0$ .

23.  $H_0$ : Proportions are as stated;  $H_1$ : Proportions are not as stated.  
 Reject  $H_0$  if  $\chi^2 > 11.345$ .

$$\chi^2 = \frac{(50 - 25)^2}{25} + \dots + \frac{(160 - 275)^2}{275} = 115.22$$

Reject  $H_0$ . The proportions are not as stated.

25.  $H_0$ : There is no relationship between community size and section read.  $H_1$ : There is a relationship.  
 Reject  $H_0$  if  $\chi^2 > 9.488$ .

$$\chi^2 = \frac{(170 - 157.50)^2}{157.50} + \dots + \frac{(88 - 83.62)^2}{83.62} = 7.340$$

Do not reject  $H_0$ . There is no relationship between community size and section read.

27.  $H_0$ : No relationship between error rates and item type.  
 $H_1$ : There is a relationship between error rates and item type.  
 Reject  $H_0$  if  $\chi^2 > 9.21$ .

$$\chi^2 = \frac{(20 - 14.1)^2}{14.1} + \dots + \frac{(225 - 225.25)^2}{225.25} = 8.033$$

Do not reject  $H_0$ . There is not a relationship between error rates and item type.

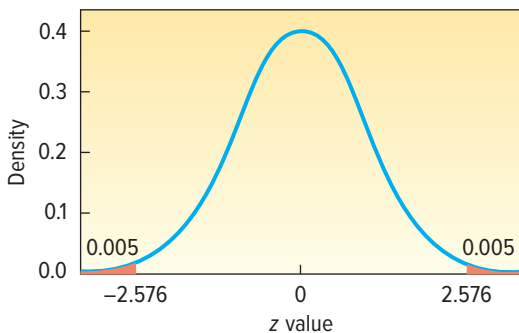
29. a. This is a binomial situation with both the mean number of successes and failures equal to 25.5, found by  $0.5(51)$ .

b.  $H_0: \pi = 0.50$      $H_1: \pi \neq 0.50$

c.

Distribution Plot

Normal, Mean = 0, StDev = 1



Reject  $H_0$  if  $z$  is not between  $-2.576$  and  $2.576$ .

d.  $z = \frac{\frac{35}{51} - 0.5}{\sqrt{0.5(1-0.5)/51}} = 2.66$

We reject the null hypothesis. These data show that the National Football Conference tends to correctly call the coin toss more often than the American Football Conference.

- e. The  $p$ -value is 0.0078, found by  $2(0.5000 - 0.4961)$ . The  $p$ -value indicates that the probability of observing a probability of .69 (35/51) or greater while assuming the null hypothesis correct (probability is 0.5) is very unlikely.

31.  $H_0: \pi \leq 0.60$      $H_1: \pi > 0.60$   
 $H_0$  is rejected if  $z > 2.33$ .

$$z = \frac{.70 - .60}{\sqrt{\frac{.60(.40)}{200}}} = 2.89$$

$H_0$  is rejected. Ms. Dennis is correct. More than 60% of the accounts are more than three months old.

33.  $H_0: \pi \leq 0.44$      $H_1: \pi > 0.44$   
 $H_0$  is rejected if  $z > 1.65$ .

$$z = \frac{0.480 - 0.44}{\sqrt{(0.44 \times 0.56)/1000}} = 2.55$$

$H_0$  is rejected. We conclude that there has been an increase in the proportion of people wanting to go to Europe.

35.  $H_0: \pi \leq 0.20$      $H_1: \pi > 0.20$   
 $H_0$  is rejected if  $z > 2.33$ .

$$z = \frac{(56/200) - 0.20}{\sqrt{(0.20 \times 0.80)/200}} = 2.83$$

$H_0$  is rejected. More than 20% of the owners move during a particular year.  $p$ -value =  $0.5000 - 0.4977 = 0.0023$ .

37.  $H_0: \pi \geq 0.0008$      $H_1: \pi < 0.0008$   
 $H_0$  is rejected if  $z < -1.645$ .

$$z = \frac{0.0006 - 0.0008}{\sqrt{\frac{0.0008(0.9992)}{10,000}}} = -0.707$$
     $H_0$  is not rejected.

These data do not prove there is a reduced fatality rate.

39.  $H_0: \pi_1 \leq \pi_2$      $H_1: \pi_1 > \pi_2$   
 If  $z > 2.33$ , reject  $H_0$ .

$$p_c = \frac{990 + 970}{1,500 + 1,600} = 0.63$$

$$z = \frac{.6600 - .60625}{\sqrt{\frac{.63(.37)}{1,500} + \frac{.63(.37)}{1,600}}} = 3.10$$

Reject the null hypothesis. We can conclude the proportion of men who believe the division is fair is greater.

41.  $H_0: \pi_1 \leq \pi_2$      $H_1: \pi_1 > \pi_2$      $H_0$  is rejected if  $z > 1.65$ .

$$p_c = \frac{.091 + .085}{2} = .088$$

$$z = \frac{0.091 - 0.085}{\sqrt{\frac{(0.088)(0.912)}{5,000} + \frac{(0.088)(0.912)}{5,000}}} = 1.059$$

$H_0$  is not rejected. There has not been an increase in the proportion calling conditions "good." The  $p$ -value is .1446, found by  $.5000 - .3554$ . The increase in the percentages will happen by chance in one out of every seven cases.

43.  $H_0: \pi_1 = \pi_2$      $H_1: \pi_1 \neq \pi_2$

$H_0$  is rejected if  $z$  is not between  $-1.96$  and  $1.96$ .

$$p_c = \frac{100 + 36}{300 + 200} = .272$$

$$z = \frac{\frac{100}{300} - \frac{36}{200}}{\sqrt{\frac{(0.272)(0.728)}{300} + \frac{(0.272)(0.728)}{200}}} = 3.775$$

$H_0$  is rejected. There is a difference in the replies of the sexes.

45.  $H_0: \pi_s = 0.50, \pi_r = \pi_o = 0.25$

$H_1$ : Distribution is not as given above.

$df = 2$ . Reject  $H_0$  if  $\chi^2 > 4.605$ .

| Turn     | $f_o$ | $f_e$ | $f_o - f_e$ | $(f_o - f_e)^2 / f_e$ |
|----------|-------|-------|-------------|-----------------------|
| Straight | 112   | 100   | 12          | 1.44                  |
| Right    | 48    | 50    | -2          | 0.08                  |
| Left     | 40    | 50    | -10         | 2.00                  |
| Total    | 200   | 200   |             | 3.52                  |

$H_0$  is not rejected. The proportions are as given in the null hypothesis.

47.  $H_0$ : There is no preference with respect to TV stations.  
 $H_1$ : There is a preference with respect to TV stations.  
 $df = 3 - 1 = 2$ .  $H_0$  is rejected if  $\chi^2 > 5.991$ .

| TV Station | $f_o$ | $f_e$ | $f_o - f_e$ | $(f_o - f_e)^2$ | $(f_o - f_e)^2 / f_e$ |
|------------|-------|-------|-------------|-----------------|-----------------------|
| WNAE       | 53    | 50    | 3           | 9               | 0.18                  |
| WRRN       | 64    | 50    | 14          | 196             | 3.92                  |
| WSPD       | 33    | 50    | -17         | 289             | 5.78                  |
|            | 150   | 150   | 0           |                 | 9.88                  |

$H_0$  is rejected. There is a preference for TV stations.

49.  $H_0$ :  $\pi_n = 0.21$ ,  $\pi_m = 0.24$ ,  $\pi_s = 0.35$ ,  $\pi_w = 0.20$   
 $H_1$ : The distribution is not as given.  
Reject  $H_0$  if  $\chi^2 > 11.345$ .

| Region    | $f_o$ | $f_e$ | $f_o - f_e$ | $(f_o - f_e)^2 / f_e$ |
|-----------|-------|-------|-------------|-----------------------|
| Northeast | 68    | 84    | -16         | 3.0476                |
| Midwest   | 104   | 96    | 8           | 0.6667                |
| South     | 155   | 140   | 15          | 1.6071                |
| West      | 73    | 80    | -7          | 0.6125                |
| Total     | 400   | 400   | 0           | 5.9339                |

$H_0$  is not rejected. The distribution of order destinations reflects the population.

51.  $H_0$ : The proportions are the same.  
 $H_1$ : The proportions are not the same.  
Reject  $H_0$  if  $\chi^2 > 16.919$ .

| $f_o$ | $f_e$ | $f_o - f_e$ | $(f_o - f_e)^2$ | $(f_o - f_e)^2 / f_e$ |
|-------|-------|-------------|-----------------|-----------------------|
| 44    | 28    | 16          | 256             | 9.143                 |
| 32    | 28    | 4           | 16              | 0.571                 |
| 23    | 28    | -5          | 25              | 0.893                 |
| 27    | 28    | -1          | 1               | 0.036                 |
| 23    | 28    | -5          | 25              | 0.893                 |
| 24    | 28    | -4          | 16              | 0.571                 |
| 31    | 28    | 3           | 9               | 0.321                 |
| 27    | 28    | -1          | 1               | 0.036                 |
| 28    | 28    | 0           | 0               | 0.000                 |
| 21    | 28    | -7          | 49              | 1.750                 |
|       |       |             |                 | 14.214                |

Do not reject  $H_0$ . The digits are evenly distributed.

53.  $H_0$ : Gender and attitude toward the deficit are not related.  
 $H_1$ : Gender and attitude toward the deficit are related.  
Reject  $H_0$  if  $\chi^2 > 5.991$ .

$$\chi^2 = \frac{(244 - 292.41)^2}{292.41} + \frac{(194 - 164.05)^2}{164.05} + \frac{(68 - 49.53)^2}{49.53} + \frac{(305 - 256.59)^2}{256.59} + \frac{(114 - 143.95)^2}{143.95} + \frac{(25 - 43.47)^2}{43.47} = 43.578$$

Since  $43.578 > 5.991$ , you reject  $H_0$ . A person's position on the deficit is influenced by his or her gender.

55.  $H_0$ : Whether a claim is filed and age are not related.  
 $H_1$ : Whether a claim is filed and age are related.  
Reject  $H_0$  if  $\chi^2 > 7.815$ .

$$\chi^2 = \frac{(170 - 203.33)^2}{203.33} + \dots + \frac{(24 - 35.67)^2}{35.67} = 53.639$$

Reject  $H_0$ . Age is related to whether a claim is filed.

57.  $H_0$ :  $\pi_{BL} = \pi_O = .23$ ,  $\pi_V = \pi_G = .15$ ,  $\pi_{BR} = \pi_R = .12$   
 $H_1$ : The proportions are not as given. Reject  $H_0$  if  $\chi^2 > 15.086$ .

| Color  | $f_o$ | $f_e$ | $(f_o - f_e)^2 / f_e$ |
|--------|-------|-------|-----------------------|
| Blue   | 12    | 16.56 | 1.256                 |
| Brown  | 14    | 8.64  | 3.325                 |
| Yellow | 13    | 10.80 | 0.448                 |
| Red    | 14    | 8.64  | 3.325                 |
| Orange | 7     | 16.56 | 5.519                 |
| Green  | 12    | 10.80 | 0.133                 |
| Total  | 72    |       | 14.006                |

Do not reject  $H_0$ . The color distribution agrees with the manufacturer's information.

59.  $H_0$ : Salary and winning are not related.  
 $H_1$ : Salary and winning are related.  
Reject  $H_0$  if  $\chi^2 > 3.841$  with 1 degree of freedom.

| Winning | Salary     |          | Total |
|---------|------------|----------|-------|
|         | Lower half | Top half |       |
| No      | 10         | 4        | 14    |
| Yes     | 5          | 11       | 16    |
| Total   | 15         | 15       |       |

$$\chi^2 = \frac{(10 - 7)^2}{7} + \frac{(4 - 7)^2}{7} + \frac{(5 - 8)^2}{8} + \frac{(11 - 8)^2}{8} = 4.82$$

Reject  $H_0$ . Conclude that salary and winning are related.

## Solutions to Practice Tests

### PRACTICE TEST—CHAPTER 1

#### Part I

1. Statistics
2. Descriptive statistics
3. Statistical inference
4. Sample
5. Population
6. Nominal
7. Ratio
8. Ordinal
9. Interval
10. Discrete
11. Nominal
12. Nominal

#### Part II

1. a. 11.1  
b. About 3 to 1  
c. 65
2. a. Ordinal  
b. 67.7%

### PRACTICE TEST—CHAPTER 2

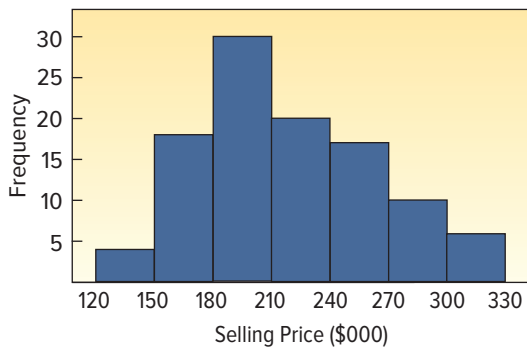
#### Part I

1. Frequency table
2. Frequency distribution
3. Bar chart
4. Pie chart
5. Histogram or frequency polygon
6. 7
7. Class interval
8. Midpoint
9. Total number of observations
10. Upper class limits

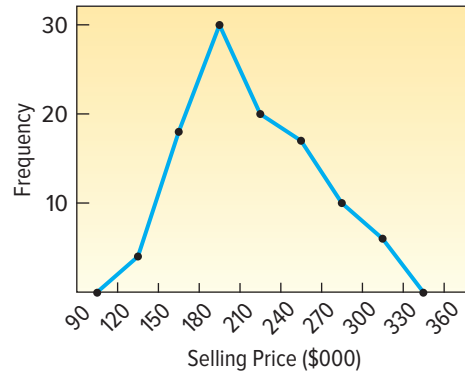
#### Part II

1. a. \$30  
b. 105  
c. 52  
d. .19  
e. \$165  
f. \$120, \$330  
g.

Selling Price of Homes in Warren, PA



h. Selling Price of Homes in Warren, PA



### PRACTICE TEST—CHAPTER 3

#### Part I

1. Parameter
2. Statistic
3. Zero
4. Median
5. 50%
6. Mode
7. Range
8. Variance
9. Variance
10. Never
11. Median
12. Normal rule or empirical rule

#### Part II

1. a.  $\bar{X} = \frac{560}{8} = 70$   
b. Median = 71.5  
c. Range =  $80 - 52 = 28$   
d.  $s = \sqrt{\frac{610.0}{8 - 1}} = 9.335$
2.  $\bar{X}_w = \frac{200(\$36) + 300(\$40) + 500(\$50)}{200 + 300 + 500} = \$44.20$
3.  $-0.88 \pm 2(1.41)$   
 $-0.88 \pm 2.82$   
 $-3.70, 1.94$

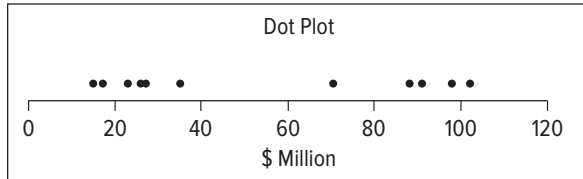
### PRACTICE TEST—CHAPTER 4

#### Part I

1. Dot plot
2. Box plot
3. Scatter diagram
4. Contingency table
5. Quartile
6. Percentile
7. Skewness
8. First quartile
9. Interquartile range

## Part II

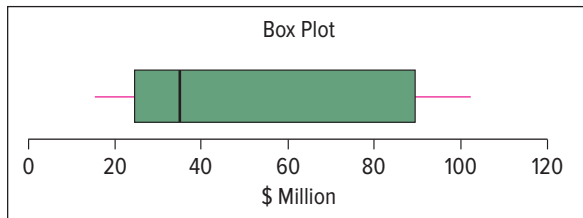
1. a.



b.  $L_{50} = (11 + 1) \frac{50}{100} = 6$   
median = 35

c.  $L_{25} = (11 + 1) \frac{25}{100} = 3$   
 $Q_1 = 23$

d.  $L_{75} = (11 + 1) \frac{75}{100} = 9$   
 $Q_3 = 91$



2. a.  $P(H) = \frac{144}{449} = 0.32$

b.  $P(H | < 30) = \frac{21}{89} = 0.24$

c.  $P(H | > 60) = \frac{75}{203} = 0.37$ . Age is related to high blood pressure, because  $P(H | > 60)$  is greater than  $P(H | < 30)$ .

## PRACTICE TEST—CHAPTER 5

### Part I

- Probability
- Experiment
- Event
- Relative frequency
- Subjective
- Classical
- Mutually exclusive
- Exhaustive
- Mutually exclusive
- Complement rule
- Joint probability
- Independent

### Part II

1. a.  $P(\text{Both}) = P(B_1) \cdot P(B_2 | B_1)$   
 $= \left(\frac{5}{20}\right) \left(\frac{4}{19}\right) = .0526$

b.  $P(\text{at least 1}) = 1 - P(\text{neither})$   
 $= 1 - \left(\frac{15}{20}\right) \left(\frac{14}{19}\right) = 1 - .5526 = .4474$

2.  $P(\text{at least 1}) = P(\text{Jogs}) + P(\text{Bike}) - P(\text{Both})$   
 $= .30 + .20 - .12 = .38$

3.  $X = 5! = 120$

## PRACTICE TEST—CHAPTER 6

### Part I

- Probability distribution
- Probability
- One
- Mean
- Two
- Never
- Equal
- $\pi$
- .075
- .183

### Part II

- a. Binomial  
b.  $P(x = 1) = {}_{16}C_1(.15)^1(.85)^{15} = (16)(.15)(.0874) = .210$   
c.  $P(x \geq 1) = 1 - P(x = 0) = 1 - {}_{16}C_0(.15)^0(.85)^{16} = .9257$
- a. Poisson  
b.  $P(x = 3) = \frac{3^3 e^{-3}}{3!} = \frac{27}{(6)(20.0855)} = .224$   
c.  $P(x = 0) = \frac{3^0 e^{-3}}{0!} = .050$   
d.  $P(x \geq 1) = 1 - P(x = 0) = 1 - .050 = .950$

3.

| Exemptions<br>x | Probability<br>P(x) | X · P(x) | (x - 2.2) <sup>2</sup> · P(x) |
|-----------------|---------------------|----------|-------------------------------|
| 1               | 0.2                 | 0.2      | 0.288                         |
| 2               | 0.5                 | 1        | 0.02                          |
| 3               | 0.2                 | 0.6      | 0.128                         |
| 4               | 0.1                 | 0.4      | 0.324                         |
|                 |                     | 2.2      | 0.76                          |

a.  $\mu = 1(.2) + 2(.5) + 3(.2) + 4(.1) = 2.2$   
b.  $\sigma^2 = (1 - 2.2)^2(.2) + \dots + (4 - 2.2)^2(.1) = 0.76$

## PRACTICE TEST—CHAPTER 7

### Part I

- One
- Infinite
- Discrete
- Always equal
- Infinite
- One
- Any of these values
- .2764
- .9396
- .0450

### Part II

1. a.  $z = \frac{2000 - 1600}{850} = .47$

$P(0 \leq z < .47) = .1808$

b.  $z = \frac{900 - 1600}{850} = -0.82$

$P(-0.82 \leq z \leq .47) = .2939 + .1808 = .4747$

c.  $z = \frac{1800 - 1600}{850} = 0.24$

$P(0.24 \leq z \leq .47) = .1808 - .0948 = .0860$

d.  $1.65 = \frac{X - 1600}{850}$

$X = 1600 + 1.65(850) = \$3002.50$



## PRACTICE TEST—CHAPTER 8

### Part I

1. Random sample
2. No size restriction
3. Strata
4. Sampling error
5. Sampling distribution of sample means
6. 120
7. Standard error of the mean
8. Always equal
9. Decrease
10. Normal distribution

### Part II

1.  $z = \frac{11 - 12.2}{2.3/\sqrt{12}} = -1.81$   
 $P(z < -1.81) = .5000 - .4649 = .0351$

## PRACTICE TEST—CHAPTER 9

### Part I

1. Point estimate
2. Confidence interval
3. Narrower
4. Proportion
5. 95
6. Standard deviation
7. Binomial
8. z distribution
9. Population median
10. Population mean

### Part II

1. a. Unknown  
 b. 9.3 years  
 c.  $s_x = \frac{2.0}{\sqrt{26}} = 0.392$   
 d.  $9.3 \pm (1.708) \frac{2.0}{\sqrt{26}}$   
 $9.3 \pm 0.67$   
 $(8.63, 9.97)$
2.  $n = (.27)(.73) \left( \frac{2.326}{.02} \right)^2 = 2.666$
3.  $.64 \pm 1.96 \sqrt{\frac{.64(.36)}{100}}$   
 $.64 \pm .094$   
 $[.546, .734]$

## PRACTICE TEST—CHAPTER 10

### Part I

1. Null hypothesis
2. Accept
3. Significant level
4. Test statistic
5. Critical value
6. Two
7. Standard deviation (or variance)
8. p-value
9. Binomial
10. Five

### Part II

1.  $H_0: \mu \leq 90, H_i: \mu > 90$   
 $df = 18 - 1 = 17$   
 Reject  $H_0$  if  $t > 2.567$ .  
 $t = \frac{96 - 90}{12/\sqrt{18}} = 2.121$

Do not reject  $H_0$ . We cannot conclude that the mean time in the park is more than 90 minutes.

2.  $H_0: \mu \leq 9.75, H_i: \mu > 9.75$

Reject  $H_0$  if  $z > 1.645$ .

Note  $\sigma$  is known, so z is used and we assume a .05 significance level.

$$z = \frac{9.85 - 9.75}{0.27/\sqrt{25}} = 1.852$$

Reject  $H_0$ . The mean weight is more than 9.75 ounces.

3.  $H_0: \pi \geq 0.67, H_i: \pi < 0.67$

Reject  $H_0$  if  $z < -1.645$ .

$$z = \frac{\frac{180}{300} - 0.67}{\sqrt{\frac{0.67(1 - 0.67)}{300}}} = -2.578$$

Reject  $H_0$ . Less than .67 of the couples seek their mate's approval.

## PRACTICE TEST—CHAPTER 11

### Part I

1. Zero
2. z
3. Proportions
4. Population standard deviation
5. Difference
6. t distribution
7.  $n - 2$
8. Paired
9. Independent
10. Dependent sample

### Part II

1.  $H_0: \mu_y = \mu_n; H_i: \mu_y \neq \mu_n$   
 $df = 14 + 12 - 2 = 24$   
 Reject  $H_0$  if  $t < -2.064$  or  $t > 2.064$ .  
 $s_p^2 = \frac{(14 - 1)30^2 + (12 - 1)(40)^2}{14 + 12 - 2} = 1220.83$   
 $t = \frac{837 - 797}{\sqrt{1220.83 \left( \frac{1}{14} + \frac{1}{12} \right)}} = \frac{40.0}{13.7455} = 2.910$   
 Reject  $H_0$ . There is a difference in the mean miles traveled.
2.  $H_0: \pi_E = \pi_T, H_i: \pi_E \neq \pi_T$   
 Reject  $H_0$  if  $z < -1.96$  or  $z > 1.96$ .  
 $P_C = \frac{128 + 149}{300 + 400} = \frac{277}{700} = .396$   
 $z = \frac{\frac{128}{300} - \frac{149}{400}}{\sqrt{\frac{.396(1 - .396)}{300} + \frac{.396(1 - .396)}{400}}} = \frac{.054}{.037} = 1.459$   
 Do not reject  $H_0$ . There is no difference on the proportion that liked the soap in the two cities.

## PRACTICE TEST—CHAPTER 12

### Part I

1. F distribution
2. Positively skewed
3. Variances
4. Means
5. Population standard deviations
6. Error or residual
7. Equal
8. Degrees of freedom
9. Variances
10. Independent

### Part II

1.  $H_0: \sigma_h^2 = \sigma_y^2; H_1: \sigma_h^2 \neq \sigma_y^2$   
 $df_y = 12 - 1 = 11$        $df_h = 14 - 1 = 13$   
 Reject  $H_0$  if  $F > 2.635$ .

$$F = \frac{(40)^2}{(30)^2} = 1.78$$

Do not reject  $H_0$ . Cannot conclude there is a difference in the variation of the miles traveled.

2. a. 3  
 b. 21  
 c. 3.55  
 d.  $H_0: \mu_1 = \mu_2 = \mu_3$   
 $H_1$ : not all treatment means are the same.  
 e. Reject  $H_0$ .  
 f. The treatment means are not the same.

### PRACTICE TEST—CHAPTER 13

#### Part I

- Scatter diagram
- 1 and 1
- Less than zero
- Coefficient of determination
- $t$
- Predicted or fitted
- Sign
- Larger
- Error
- Independent

#### Part II

1. a. 25  
 b. Shares of stock  
 c.  $\hat{Y} = 197.9229 + 24.9145X$   
 d. Direct  
 e.  $r = \sqrt{\frac{152,399.0211}{208,333.1400}} = 0.855$   
 f.  $\hat{Y} = 197.9229 + 24.9145(10) = 447.0679$ , or 447  
 g. Increase almost 25  
 h.  $H_0: \beta \leq 0$   
 $H_1: \beta > 0$   
 Reject  $H_0$  if  $t > 1.71$   
 $t = \frac{24.9145}{3.1473} = 7.916$   
 Reject  $H_0$ . There is a positive relationship between years and shares.

### PRACTICE TEST—CHAPTER 14

#### Part I

- |                          |                       |
|--------------------------|-----------------------|
| 1. Independent variables | 7. $F$ distribution   |
| 2. Least squares         | 8. $t$ distribution   |
| 3. Mean square error     | 9. Linearity          |
| 4. Independent variables | 10. Correlated        |
| 5. Independent variable  | 11. Multicollinearity |
| 6. Different from zero   | 12. Dummy variable    |

### Part II

1. a. Four  
 b.  $\hat{Y} = 70.06 + 0.42x_1 + 0.27x_2 + 0.75x_3 + 0.42x_4$   
 c.  $R^2 = \frac{1050.8}{1134.6} = 0.926$   
 d.  $s_{y,1234} = \sqrt{4.19} = 2.05$   
 e.  $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$   
 $H_1$ : not all  $\beta_i = 0$   
 Reject  $H_0$  if  $F > 2.87$ .  
 $F = \frac{262.70}{4.19} = 62.70$   
 Reject  $H_0$ . Not all the regression coefficients equal zero.  
 d.  $H_0: \beta_i = 0, H_1: \beta_i \neq 0$   
 Reject  $H_0$  if  $t < -2.086$  or  $t > 2.086$ .

|                  |                     |                  |                  |
|------------------|---------------------|------------------|------------------|
| $\beta_1 = 0$    | $\beta_2 = 0$       | $\beta_3 = 0$    | $\beta_4 = 0$    |
| $\beta_1 \neq 0$ | $\beta_2 \neq 0$    | $\beta_3 \neq 0$ | $\beta_4 \neq 0$ |
| $t = 2.47$       | $t = 1.29$          | $t = 2.50$       | $t = 6.00$       |
| Reject $H_0$     | Do not reject $H_0$ | Reject $H_0$     | Reject $H_0$     |

Conclusion. Drop variable 2 and retain the others.

### PRACTICE TEST—CHAPTER 15

#### Part I

- Nominal
- No assumption
- Can have negative values
- 2
- 6
- Independent
- 4
- The same
- 9.488
- Degrees of freedom

#### Part II

1.  $H_0$ : There is no difference between the school district and census data.  
 $H_1$ : There is a difference between the school district and census data.  
 Reject  $H_0$  if  $\chi^2 > 7.815$ .  

$$\chi^2 = \frac{(120 - 130)^2}{130} + \frac{(40 - 40)^2}{40} + \frac{(30 - 20)^2}{20} + \frac{(10 - 10)^2}{10} = 5.77$$
 Do not reject  $H_0$ . There is no difference between the census and school district data.
2.  $H_0$ : Gender and book type are independent.  
 $H_1$ : Gender and book type are related.  
 Reject  $H_0$  if  $\chi^2 > 5.991$ .  

$$\chi^2 = \frac{(250 - 197.31)^2}{197.31} + \dots + \frac{(200 - 187.5)^2}{187.5} = 54.842$$
 Reject  $H_0$ . Men and women read different types of books.

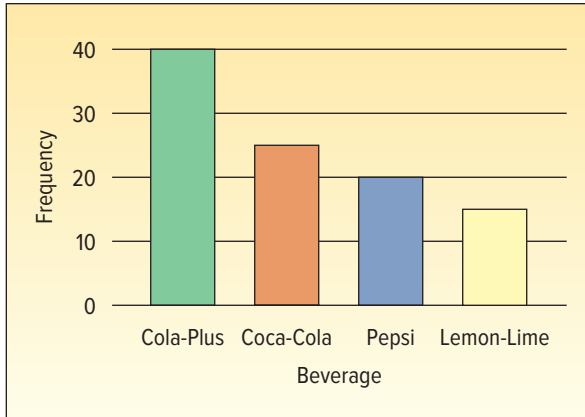
# APPENDIX E: ANSWERS TO SELF-REVIEW

## CHAPTER 1

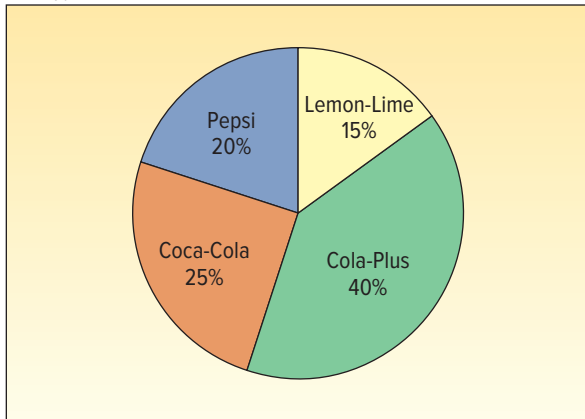
- 1-1** a. Inferential statistics, because a sample was used to draw a conclusion about how all consumers in the population would react if the chicken dinner were marketed.  
 b. On the basis of the sample of 1,960 consumers, we estimate that, if it is marketed, 60% of all consumers will purchase the chicken dinner:  $(1,176/1,960) \times 100 = 60\%$ .
- 1-2** a. Age is a ratio-scale variable. A 40-year-old is twice as old as someone 20 years old.  
 b. The two variables are: 1) if a person owns a luxury car, and 2) the state of residence. Both are measured on a nominal scale.

## CHAPTER 2

- 2-1** a. Qualitative data, because the customers' response to the taste test is the name of a beverage.  
 b. Frequency table. It shows the number of people who prefer each beverage.  
 c.



d.



- 2-2** a. The raw data or ungrouped data.  
 b.

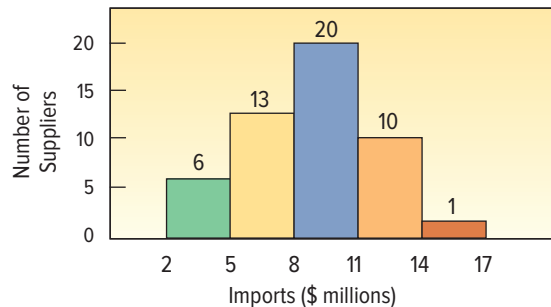
| Commission            | Number of Salespeople |
|-----------------------|-----------------------|
| \$1,400 up to \$1,500 | 2                     |
| 1,500 up to 1,600     | 5                     |
| 1,600 up to 1,700     | 3                     |
| 1,700 up to 1,800     | 1                     |
| Total                 | 11                    |

- c. Class frequencies.  
 d. The largest concentration of commissions is \$1,500 up to \$1,600. The smallest commission is about \$1,400 and the largest is about \$1,800. The typical amount earned is \$1,550.
- 2-3** a.  $2^6 = 64 < 73 < 128 = 2^7$ , so seven classes are recommended.  
 b. The interval width should be at least  $(488 - 320)/7 = 24$ . Class intervals of either 25 or 30 are reasonable.  
 c. Assuming a class interval of 25 and beginning with a lower limit of 300, eight classes are required. If we use an interval of 30 and begin with a lower limit of 300, only seven classes are required. Seven classes is the better alternative.

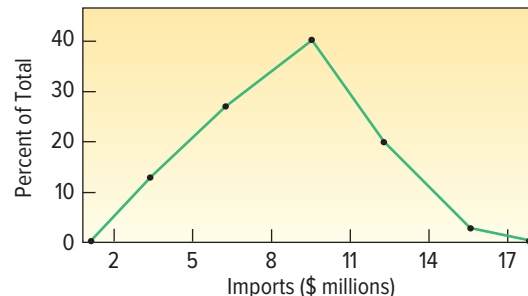
| Distance Classes | Frequency | Percent |
|------------------|-----------|---------|
| 300 up to 330    | 2         | 2.7%    |
| 330 up to 360    | 2         | 2.7     |
| 360 up to 390    | 17        | 23.3    |
| 390 up to 420    | 27        | 37.0    |
| 420 up to 450    | 22        | 30.1    |
| 450 up to 480    | 1         | 1.4     |
| 480 up to 510    | 2         | 2.7     |
| Grand Total      | 73        | 100.00  |

- d. 17  
 e. 23.3%, found by  $17/73$   
 f. 71.2%, found by  $(27 + 22 + 1 + 2)/73$

**2-4** a.



b.

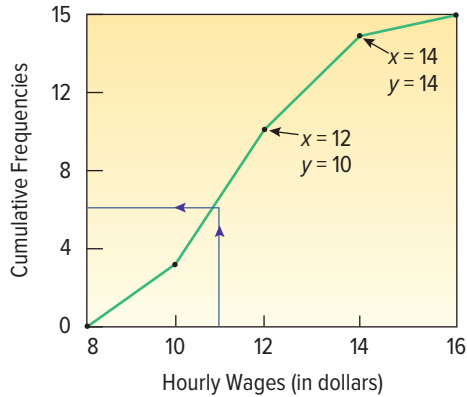


The plots are: (3.5, 12), (6.5, 26), (9.5, 40), (12.5, 20), and (15.5, 2).

- c. The smallest annual volume of imports by a supplier is about \$2 million, the largest about \$17 million. The highest frequency is between \$8 million and \$11 million.

**2-5** a. A frequency distribution.

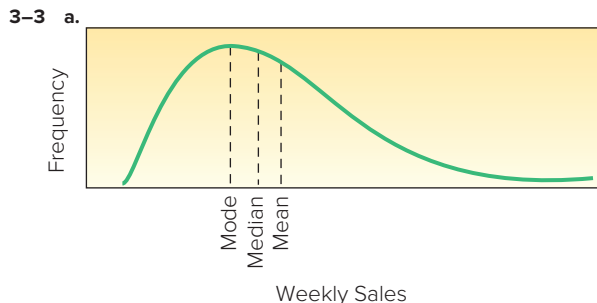
| Hourly Wages   | Cumulative Number |
|----------------|-------------------|
| Less than \$8  | 0                 |
| Less than \$10 | 3                 |
| Less than \$12 | 10                |
| Less than \$14 | 14                |
| Less than \$16 | 15                |



c. About seven employees earn \$11.00 or less.

### CHAPTER 3

- 3-1** a.  $\bar{x} = \frac{\sum x}{n}$   
 b.  $\bar{x} = \frac{\$267,100}{4} = \$66,775$   
 c. Statistic, because it is a sample value.  
 d. \$66,775. The sample mean is our best estimate of the population mean.
- 2.** a.  $\mu = \frac{\sum x}{N}$   
 b.  $\mu = \frac{498}{6} = 83$   
 c. Parameter, because it was computed using all the population values.
- 3-2** 1. a. \$878  
 b. 3, 3  
 2. a. 17, found by  $(15 + 19)/2 = 17$   
 b. 5, 5  
 c. There are 3 values that occur twice: 11, 15, and 19. There are three modes.



- b. Positively skewed, because the mean is the largest average and the mode is the smallest.
- 3-4** a. \$237, found by:  

$$\frac{(95 \times \$400) + (126 \times \$200) + (79 \times \$100)}{95 + 126 + 79} = \$237.00$$
  
 b. The profit per suit is \$12, found by  $\$237 - \$200$  cost - \$25 commission. The total profit for the 300 suits is \$3,600, found by  $300 \times \$12$ .

- 3-5** a. 22 thousand pounds, found by  $112 - 90$   
 b.  $\bar{x} = \frac{824}{8} = 103$  thousand pounds  
 c. Variance =  $\frac{373}{8} = 46.625$
- 3-6** a.  $\mu = \frac{\$16,900}{5} = \$3,380$   
 b.  $\sigma^2 = \frac{(3,536 - 3,380)^2 + \dots + (3,622 - 3,380)^2}{5}$   

$$= \frac{(156)^2 + (-207)^2 + (68)^2 + (-259)^2 + (242)^2}{5}$$
  

$$= \frac{197,454}{5} = 39,490.8$$
  
 c.  $\sigma = \sqrt{39,490.8} = 198.72$   
 d. There is more variation in the Pittsburgh office because the standard deviation is larger. The mean is also larger in the Pittsburgh office.
- 3-7** 2.33, found by:  

$$\bar{x} = \frac{\sum x}{n} = \frac{28}{7} = 4$$
  

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$
  

$$= \frac{14}{7 - 1}$$
  

$$= 2.33$$
  

$$s = \sqrt{2.33} = 1.53$$
- 3-8** a.  $k = \frac{14.15 - 14.00}{.10} = 1.5$   

$$k = \frac{13.85 - 14.0}{.10} = -1.5$$
  

$$1 - \frac{1}{(1.5)^2} = 1 - .44 = .56$$
  
 b. 13.8 and 14.2

### CHAPTER 4

- 4-1** a. 79, 105  
 b. 15  
 c. From 88 to 97; 75% of the stores are in this range.
- 4-2** a. 7.9  
 b.  $Q_1 = 7.76$ ,  $Q_3 = 8.015$
- 4-3** The smallest value is 10 and the largest 85; the first quartile is 25 and the third 60. About 50% of the values are between 25 and 60. The median value is 40. The distribution is positively skewed. There are no outliers.
- 4-4** a.  $\bar{x} = \frac{407}{5} = 81.4$   

$$s = \sqrt{\frac{923.2}{5 - 1}} = 15.19$$
, Median = 84  
 b.  $sk = \frac{3(81.4 - 84.0)}{15.19} = -0.51$   
 c.  $sk = \frac{5}{(4)(3)} [-1.3154] = -0.5481$   
 d. The distribution is somewhat negatively skewed.
- 4-5** a. Scatter diagram  
 b. 16  
 c. \$7,500  
 d. Strong and direct

### CHAPTER 5

- 5-1** a. Count the number who think the new game is playable.  
 b. Seventy-three players found the game playable. Many other answers are possible.

- c. No. Probability cannot be greater than 1. The probability that the game, if put on the market, will be successful is  $65/80$ , or  $.8125$ .
- d. Cannot be less than 0. Perhaps a mistake in arithmetic.
- e. More than half of the players testing the game liked it. (Of course, other answers are possible.)

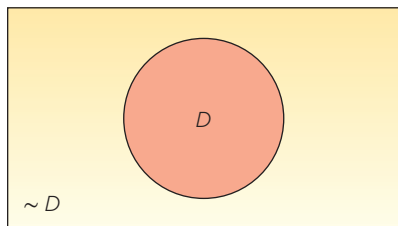
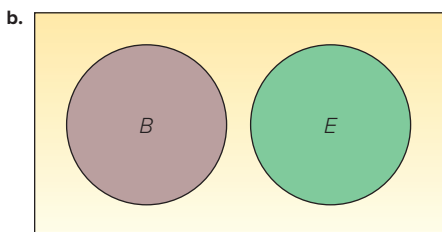
5-2 1.  $\frac{4 \text{ queens in deck}}{52 \text{ cards total}} = \frac{4}{52} = .0769$  Classical.

2.  $\frac{182}{539} = .338$  Empirical.

3. The probability of the outcome is estimated by applying the subjective approach to estimating a probability. If you think that it is likely that you will save \$1 million, then your probability should be between  $.5$  and  $1.0$ .

5-3 a. i.  $\frac{(50 + 68)}{2,000} = .059$

ii.  $1 - \frac{302}{2,000} = .849$



c. They are not complementary, but are mutually exclusive.

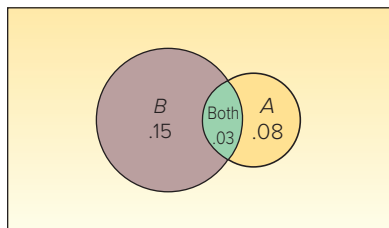
- 5-4 a. Need for corrective shoes is event  $A$ . Need for major dental work is event  $B$ .

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

$$= .08 + .15 - .03$$

$$= .20$$

b. One possibility is:



5-5  $(.95)(.95)(.95)(.95) = .8145$

5-6 a.  $.002$ , found by:

$$\left(\frac{4}{12}\right)\left(\frac{3}{11}\right)\left(\frac{2}{10}\right)\left(\frac{1}{9}\right) = \frac{24}{11,880} = .002$$

b.  $.14$ , found by:

$$\left(\frac{8}{12}\right)\left(\frac{7}{11}\right)\left(\frac{6}{10}\right)\left(\frac{5}{9}\right) = \frac{1,680}{11,880} = .1414$$

c. No, because there are other possibilities, such as three women and one man.

5-7 a.  $P(B_2) = \frac{225}{500} = .45$

b. The two events are mutually exclusive, so apply the special rule of addition.

$$P(B_1 \text{ or } B_2) = P(B_1) + P(B_2) = \frac{100}{500} + \frac{225}{500} = .65$$

c. The two events are not mutually exclusive, so apply the general rule of addition.

$$P(B_1 \text{ or } A_1) = P(B_1) + P(A_1) - P(B_1 \text{ and } A_1)$$

$$= \frac{100}{500} + \frac{75}{500} - \frac{15}{500} = .32$$

d. As shown in the example/solution, movies attended per month and age are not independent, so apply the general rule of multiplication.

$$P(B_1 \text{ and } A_1) = P(B_1)P(A_1 | B_1)$$

$$= \left(\frac{100}{500}\right)\left(\frac{15}{100}\right) = .03$$

5-8 a.  $P(\text{visited often}) = \frac{80}{195} = .41$

b.  $P(\text{visited a store in an enclosed mall}) = \frac{90}{195} = .46$

c. The two events are not mutually exclusive, so apply the general rule of addition.

$$P(\text{visited often or visited a Sears in an enclosed mall})$$

$$= P(\text{often}) + P(\text{enclosed mall}) - P(\text{often and enclosed mall})$$

$$= \frac{80}{195} + \frac{90}{195} - \frac{60}{195} = .67$$

d.  $P(\text{visited often} | \text{went to a Sears in an enclosed mall})$

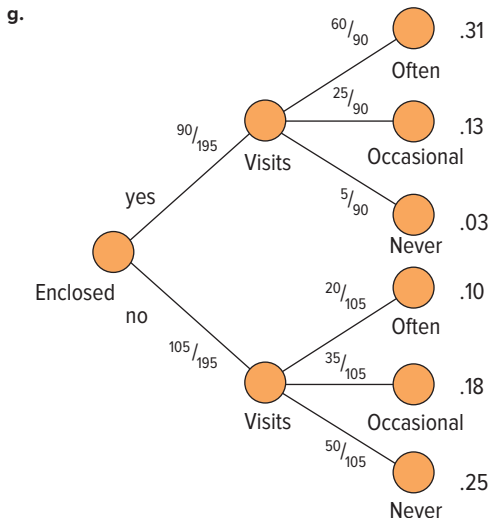
$$= \frac{60}{90} = .67$$

e. Independence requires that  $P(A|B) = P(A)$ . One possibility is:  $P(\text{visit often} | \text{visited an enclosed mall}) = P(\text{visit often})$ . Does  $60/90 = 80/195$ ? No, the two variables are not independent. Therefore, any joint probability in the table must be computed by using the general rule of multiplication.

f. As shown in part (e), visits often and enclosed mall are not independent, so apply the general rule of multiplication.

$$P(\text{often and enclosed mall}) = P(\text{often})P(\text{enclosed} | \text{often})$$

$$= \left(\frac{80}{195}\right)\left(\frac{60}{80}\right) = .31$$



- 5-9 1.  $(5)(4) = 20$   
 2.  $(3)(2)(4)(3) = 72$
- 5-10 1. a. 60, found by  $(5)(4)(3)$ .  
 b. 60, found by:  

$$\frac{5!}{(5-3)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1}$$
2. 5,040, found by:  

$$\frac{10!}{(10-4)!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}$$
3. a. 35 is correct, found by:  

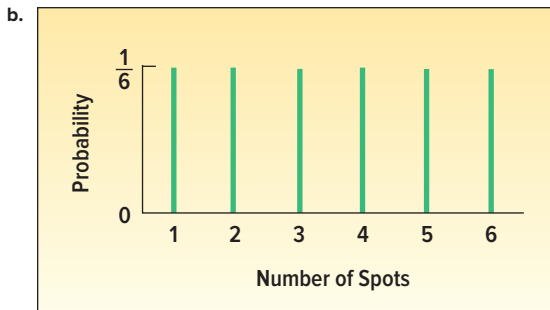
$${}_7C_3 = \frac{n!}{r!(n-r)!} = \frac{7!}{3!(7-3)!} = 35$$
  
 b. Yes. There are 21 combinations, found by:  

$${}_7C_5 = \frac{n!}{r!(n-r)!} = \frac{7!}{5!(7-5)!} = 21$$
4. a.  ${}_{50}P_3 = \frac{50!}{(50-3)!} = 117,600$   
 b.  ${}_{50}C_3 = \frac{50!}{3!(50-3)!} = 19,600$

## CHAPTER 6

6-1 a.

| Number of Spots | Probability          |
|-----------------|----------------------|
| 1               | $\frac{1}{6}$        |
| 2               | $\frac{1}{6}$        |
| 3               | $\frac{1}{6}$        |
| 4               | $\frac{1}{6}$        |
| 5               | $\frac{1}{6}$        |
| 6               | $\frac{1}{6}$        |
| Total           | $\frac{6}{6} = 1.00$ |



- c.  $\frac{6}{6}$  or 1.
- 6-2 a. It is discrete because the values \$1.99, \$2.49, and \$2.89 are clearly separated from each other. Also the sum of the probabilities is 1.00, and the outcomes are mutually exclusive.

b.

| $x$  | $P(x)$ | $xP(x)$     |
|------|--------|-------------|
| 1.99 | .30    | 0.597       |
| 2.49 | .50    | 1.245       |
| 2.89 | .20    | 0.578       |
|      |        | Sum is 2.42 |

Mean is 2.42

c.

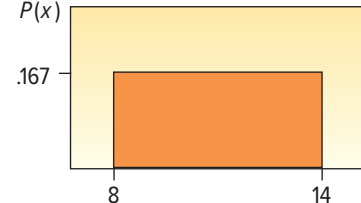
| $x$  | $P(x)$ | $(x - \mu)$ | $(x - \mu)^2 P(x)$ |
|------|--------|-------------|--------------------|
| 1.99 | .30    | -0.43       | 0.05547            |
| 2.49 | .50    | 0.07        | 0.00245            |
| 2.89 | .20    | 0.47        | 0.04418            |
|      |        |             | 0.10210            |

The variance is 0.10208, and the standard deviation is 31.95 cents.

- 6-3 a. It is reasonable because each employee either uses direct deposit or does not; employees are independent; the probability of using direct deposit is 0.95 for all; and we count the number using the service out of 7.  
 b.  $P(7) = {}_7C_7 (.95)^7 (.05)^0 = .6983$   
 c.  $P(4) = {}_7C_4 (.95)^4 (.05)^3 = .0036$   
 d. Answers are in agreement.
- 6-4 a.  $n = 8, \pi = .40$   
 b.  $P(x = 3) = .2787$   
 c.  $P(x > 0) = 1 - P(x = 0) = 1 - .0168 = .9832$
- 6-5  $\mu = 4,000(.0002) = 0.8$   
 $P(1) = \frac{0.8^1 e^{-0.8}}{1!} = .3595$

## CHAPTER 7

7-1 a.



- b.  $P(x) = (\text{height})(\text{base})$   

$$= \left(\frac{1}{14-8}\right)(14-8)$$
  

$$= \left(\frac{1}{6}\right)(6) = 1.00$$
- c.  $\mu = \frac{a+b}{2} = \frac{14+8}{2} = \frac{22}{2} = 11$   

$$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(14-8)^2}{12}} = \sqrt{\frac{36}{12}} = \sqrt{3}$$
  

$$= 1.73$$
- d.  $P(10 < x < 14) = (\text{height})(\text{base})$   

$$= \left(\frac{1}{14-8}\right)(14-10)$$
  

$$= \frac{1}{6}(4)$$
  

$$= .667$$
- e.  $P(x < 9) = (\text{height})(\text{base})$   

$$= \left(\frac{1}{14-8}\right)(9-8)$$
  

$$= 0.167$$

- 7-2 a.  $z = (64 - 48)/12.8 = 1.25$ . This person's difference of 16 ounces more than average is 1.25 standard deviations above the average.  
 b.  $z = (32 - 48)/12.8 = -1.25$ . This person's difference of 16 ounces less than average is 1.25 standard deviations below the average.
- 7-3 a. \$46,400 and \$48,000, found by  $\$47,200 \pm 1(\$800)$ .  
 b. \$45,600 and \$48,800, found by  $\$47,200 \pm 2(\$800)$ .  
 c. \$44,800 and \$49,600, found by  $\$47,200 \pm 3(\$800)$ .  
 d. \$47,200. The mean, median, and mode are equal for a normal distribution.  
 e. Yes, a normal distribution is symmetrical.

7-4 a. Computing z:

$$z = \frac{154 - 150}{5} = 0.80$$

Referring to Appendix B.3, the area is .2881. So  $P(150 < \text{temp} < 154) = .2881$ .

b. Computing z:

$$z = \frac{164 - 150}{5} = 2.80$$

Referring to Appendix B.3, the area is .4974. So  $P(164 > \text{temp}) = .5000 - .4974 = .0026$

7-5 a. Computing the z values:

$$z = \frac{146 - 150}{5} = -0.80 \quad \text{and} \quad z = \frac{156 - 150}{5} = 1.20$$

$$P(146 < \text{temp} < 156) = P(-0.80 < z < 1.20) = .2881 + .3849 = .6730$$

b. Computing the z values:

$$z = \frac{162 - 150}{5} = 2.40 \quad \text{and} \quad z = \frac{156 - 150}{5} = 1.20$$

$$P(156 < \text{temp} < 162) = P(1.20 < z < 2.40) = .4918 - .3849 = .1069$$

7-6 85.24 (instructor would no doubt make it 85). The closest area to .4000 is .3997; z is 1.28. Then:

$$1.28 = \frac{x - 75}{8}$$

$$10.24 = x - 75$$

$$x = 85.24$$

## CHAPTER 8

8-1 a. Students selected are Price, Detley, and Molter.

b. Answers will vary.

c. Skip it and move to the next random number.

8-2 The students selected are Berry, Francis, Kopp, Poteau, and Swetye.

8-3 a. 10, found by:

$${}_5C_2 = \frac{5!}{2!(5-2)!}$$

|    | Service       | Sample Mean |
|----|---------------|-------------|
| b. | Snow, Tolson  | 21          |
|    | Snow, Kraft   | 23          |
|    | Snow, Irwin   | 22          |
|    | Snow, Jones   | 24          |
|    | Tolson, Kraft | 24          |
|    | Tolson, Irwin | 23          |
|    | Tolson, Jones | 25          |
|    | Kraft, Irwin  | 25          |
|    | Kraft, Jones  | 27          |
|    | Irwin, Jones  | 26          |

| Mean | Number | Probability |
|------|--------|-------------|
| 21   | 1      | .10         |
| 22   | 1      | .10         |
| 23   | 2      | .20         |
| 24   | 2      | .20         |
| 25   | 2      | .20         |
| 26   | 1      | .10         |
| 27   | 1      | .10         |
|      | 10     | 1.00        |

d. Identical: population mean,  $\mu$ , is 24, and mean of sample means is also 24.

e. Sample means range from 21 to 27. Population values go from 20 to 28.

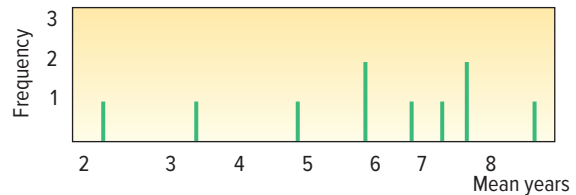
f. No, the population is uniformly distributed.

g. Yes.

8-4 The answers will vary. Here is one solution.

|           | Sample Number |          |          |          |           |           |           |           |          |          |
|-----------|---------------|----------|----------|----------|-----------|-----------|-----------|-----------|----------|----------|
|           | 1             | 2        | 3        | 4        | 5         | 6         | 7         | 8         | 9        | 10       |
|           | 8             | 2        | 2        | 19       | 3         | 4         | 0         | 4         | 1        | 2        |
|           | 19            | 1        | 14       | 9        | 2         | 5         | 8         | 2         | 14       | 4        |
|           | 8             | 3        | 4        | 2        | 4         | 4         | 1         | 14        | 4        | 1        |
|           | 0             | 3        | 2        | 3        | 1         | 2         | 16        | 1         | 2        | 3        |
|           | <u>2</u>      | <u>1</u> | <u>7</u> | <u>2</u> | <u>19</u> | <u>18</u> | <u>18</u> | <u>16</u> | <u>3</u> | <u>7</u> |
| Total     | 37            | 10       | 29       | 35       | 29        | 33        | 43        | 37        | 24       | 17       |
| $\bar{x}$ | 7.4           | 2        | 5.8      | 7.0      | 5.8       | 6.6       | 8.6       | 7.4       | 4.8      | 3.4      |

Mean of the 10 sample means is 5.88.



$$8-5 \quad z = \frac{31.08 - 31.20}{0.4/\sqrt{16}} = -1.20$$

The probability that z is greater than -1.20 is  $.5000 + .3849 = .8849$ . There is more than an 88% chance the filling operation will produce bottles with at least 31.08 ounces.

## CHAPTER 9

9-1 a. Unknown. This is the value we wish to estimate.

b. The sample mean of \$20,000 is the point estimate of the population mean daily franchise sales.

$$c. \quad \$20,000 \pm 1.960 \frac{\$3,000}{\sqrt{40}} = \$20,000 \pm \$930$$

d. The estimate of the population mean daily sales for the Bun-and-Run franchises is between \$19,070 and \$20,930. About 95% of all possible samples of 40 Bun-and-Run franchises would include the population mean.

$$9-2 \quad a. \quad \bar{x} = \frac{18}{10} = 1.8 \quad s = \sqrt{\frac{11.6}{10-1}} = 1.1353$$

b. The population mean is not known. The best estimate is the sample mean, 1.8 days.

$$c. \quad 1.80 \pm 2.262 \frac{1.1353}{\sqrt{10}} = 1.80 \pm 0.81$$

The endpoints are 0.99 and 2.61.

d. t is used because the population standard deviation is unknown.

e. The value of 0 is not in the interval. It is unreasonable to conclude that the mean number of days of work missed is 0 per employee.

$$9-3 \quad a. \quad p = \frac{420}{1,400} = .30$$

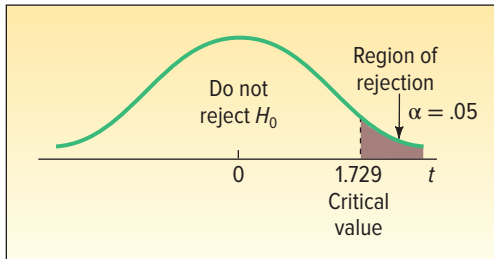
$$b. \quad .30 \pm 2.576(.0122) = .30 \pm .03$$

c. The interval is between .27 and .33. About 99% of the similarly constructed intervals would include the population mean.

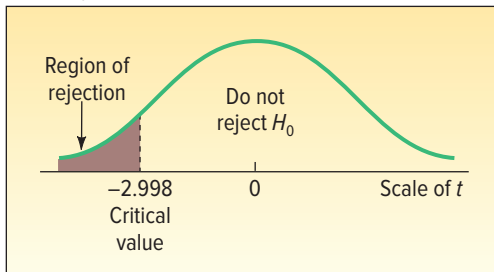
$$9-4 \quad n = \left( \frac{2.576(.279)}{.05} \right)^2 = 206.6. \quad \text{The sample should be rounded to 207.}$$

## CHAPTER 10

- 10-1 a.  $H_0: \mu = 16.0; H_1: \mu \neq 16.0$   
 b. .05  
 c.  $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$   
 d. Reject  $H_0$  if  $z < -1.96$  or  $z > 1.96$ .  
 e.  $z = \frac{16.017 - 16.0}{0.15/\sqrt{50}} = \frac{0.0170}{0.0212} = 0.80$   
 f. Do not reject  $H_0$ .  
 g. We cannot conclude the mean amount dispensed is different from 16.0 ounces.
- 10-2 a.  $H_0: \mu \leq 16.0; H_1: \mu > 16.0$   
 b. Reject  $H_0$  if  $z > 1.645$ .  
 c.  $z = \frac{16.040 - 16.0}{0.15/\sqrt{50}} = \frac{.0400}{.0212} = 1.89$   
 d. Reject  $H_0$ .  
 e. The mean amount dispensed is more than 16.0 ounces.  
 f.  $p$ -value =  $.5000 - .4706 = .0294$ . The  $p$ -value is less than  $\alpha$  (.05), so  $H_0$  is rejected. It is the same conclusion as in part (d).
- 10-3 a.  $H_0: \mu \leq 305; H_1: \mu > 305$   
 b.  $df = n - 1 = 20 - 1 = 19$   
 The decision rule is to reject  $H_0$  if  $t > 1.729$ .



- c.  $t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{311 - 305}{12/\sqrt{20}} = 2.236$   
 Reject  $H_0$  because  $2.236 > 1.729$ . The modification increased the mean battery life to more than 305 days.
- 10-4 a.  $H_0: \mu \geq 9.0; H_1: \mu < 9.0$   
 b. 7, found by  $n - 1 = 8 - 1 = 7$   
 c. Reject  $H_0$  if  $t < -2.998$ .



- d.  $t = -2.494$ , found by:
- $$s = \sqrt{\frac{0.36}{8-1}} = 0.2268$$
- $$\bar{x} = \frac{70.4}{8} = 8.8$$

Then

$$t = \frac{8.8 - 9.0}{0.2268/\sqrt{8}} = -2.494$$

Since  $-2.494$  lies to the right of  $-2.998$ ,  $H_0$  is not rejected. We have not shown that the mean is less than 9.0.

- e. The  $p$ -value is between .025 and .010.

## CHAPTER 11

- 11-1 a.  $H_0: \mu_W \leq \mu_M; H_1: \mu_W > \mu_M$   
 The subscript  $W$  refers to the women and  $M$  to the men.  
 b. Reject  $H_0$  if  $z > 1.645$ .  
 c.  $z = \frac{\$1,500 - \$1,400}{\sqrt{\frac{(\$250)^2}{50} + \frac{(\$200)^2}{40}}} = 2.11$   
 d. Reject the null hypothesis.  
 e.  $p$ -value =  $.5000 - .4826 = .0174$   
 f. The mean amount sold per day is larger for women.
- 11-2 a.  $H_0: \mu_d = \mu_o; H_1: \mu_d \neq \mu_o$   
 b.  $df = 6 + 8 - 2 = 12$   
 Reject  $H_0$  if  $t < -2.179$  or  $t > 2.179$ .  
 c.  $\bar{x}_1 = \frac{42}{6} = 7.00; s_1 = \sqrt{\frac{10}{6-1}} = 1.4142$   
 $\bar{x}_2 = \frac{80}{8} = 10.00; s_2 = \sqrt{\frac{36}{8-1}} = 2.2678$   
 $s_p^2 = \frac{(6-1)(1.4142)^2 + (8-1)(2.2678)^2}{6+8-2} = 3.8333$   
 $t = \frac{7.00 - 10.00}{\sqrt{3.8333\left(\frac{1}{6} + \frac{1}{8}\right)}} = -2.837$   
 d. Reject  $H_0$  because  $-2.837$  is less than the critical value.  
 e. The  $p$ -value is less than .02.  
 f. The mean number of defects is not the same on the two shifts.  
 g. Independent populations, populations follow the normal distribution, populations have equal standard deviations.
- 11-3 a.  $H_0: \mu_d \geq 0; H_1: \mu_d < 0$   
 b. Reject  $H_0$  if  $t > 2.998$ .  
 c.

| Name     | Before | After | $d$ | $(d - \bar{d})$ | $(d - \bar{d})^2$ |
|----------|--------|-------|-----|-----------------|-------------------|
| Hunter   | 155    | 154   | 1   | -7.875          | 62.0156           |
| Cashman  | 228    | 207   | 21  | 12.125          | 147.0156          |
| Mervine  | 141    | 147   | -6  | -14.875         | 221.2656          |
| Massa    | 162    | 157   | 5   | -3.875          | 15.0156           |
| Creola   | 211    | 196   | 15  | 6.125           | 37.5156           |
| Peterson | 164    | 150   | 14  | 5.125           | 26.2656           |
| Redding  | 184    | 170   | 14  | 5.125           | 26.2656           |
| Poust    | 172    | 165   | 7   | -1.875          | 3.5156            |
|          |        |       | 71  |                 | 538.8750          |

$$\bar{d} = \frac{71}{8} = 8.875$$

$$s_d = \sqrt{\frac{538.875}{8-1}} = 8.774$$

$$t = \frac{8.875}{8.774/\sqrt{8}} = 2.861$$

- d. Do not reject  $H_0$ . We cannot conclude that the students lost weight. The  $p$ -value is less than .025 but larger than .01.  
 e. The distribution of the differences must be approximately normal.

## CHAPTER 12

- 12-1 Let Mark's assemblies be population 1, then  $H_0: \sigma_1^2 \leq \sigma_2^2; H_1: \sigma_1^2 > \sigma_2^2; df_1 = 10 - 1 = 9$ ; and  $df_2$  also equals 9.  $H_0$  is rejected if  $F > 3.18$ .

$$F = \frac{(2.0)^2}{(1.5)^2} = 1.78$$

$H_0$  is not rejected. The variation is the same for both employees.

- 12-2 a.  $H_0: \mu_1 = \mu_2 = \mu_3$   
 $H_1$ : At least one treatment mean is different.



b. Reject  $H_0$  if  $F > 4.26$ .

c.  $\bar{x} = \frac{240}{12} = 20$

$$SS \text{ total} = (18 - 20)^2 + \dots + (32 - 20)^2 = 578$$

$$SSE = (18 - 17)^2 + (14 - 17)^2 + \dots + (32 - 29)^2 = 74$$

$$SST = 578 - 74 = 504$$

d.

| Source    | Sum of Squares | Degrees of Freedom | Mean Square | F     |
|-----------|----------------|--------------------|-------------|-------|
| Treatment | 504            | 2                  | 252         | 30.65 |
| Error     | 74             | 9                  | 8.22        |       |
| Total     | 578            | 11                 |             |       |

e.  $H_0$  is rejected. There is a difference in the mean number of bottles sold at the various locations.

12-3 a.  $H_0: \mu_1 = \mu_2 = \mu_3$   
 $H_1$ : Not all means are equal.

b.  $H_0$  is rejected if  $F > 3.98$ .

c.

| ANOVA: Single Factor |       |     |         |          |
|----------------------|-------|-----|---------|----------|
| Groups               | Count | Sum | Average | Variance |
| Northeast            | 5     | 205 | 41      | 1        |
| Southeast            | 4     | 155 | 38.75   | 0.916667 |
| West                 | 5     | 184 | 36.8    | 0.7      |

| ANOVA               |          |    |          |          |          |
|---------------------|----------|----|----------|----------|----------|
| Source of Variation | SS       | df | MS       | F        | P-value  |
| Between Groups      | 44.16429 | 2  | 22.08214 | 25.43493 | 7.49E-05 |
| Within Groups       | 9.55     | 11 | 0.868182 |          |          |
| Total               | 53.71429 | 13 |          |          |          |

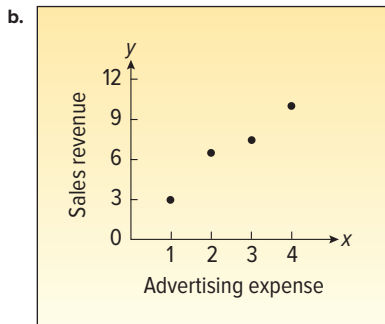
d.  $H_0$  is rejected. The treatment means differ.

e.  $(41 - 36.8) \pm 2.201 \sqrt{0.8682 \left( \frac{1}{5} + \frac{1}{5} \right)} = 4.2 \pm 1.3 = 2.9$   
 and 5.50

These treatment means differ because both endpoints of the confidence interval are of the same sign. Zero is not in the interval.

### CHAPTER 13

13-1 a. Advertising expense is the independent variable, and sales revenue is the dependent variable.



c.

| x  | y  | (x - $\bar{x}$ ) | (x - $\bar{x}$ ) <sup>2</sup> | (y - $\bar{y}$ ) | (y - $\bar{y}$ ) <sup>2</sup> | (x - $\bar{x}$ )(y - $\bar{y}$ ) |
|----|----|------------------|-------------------------------|------------------|-------------------------------|----------------------------------|
| 2  | 7  | -0.5             | .25                           | 0                | 0                             | 0                                |
| 1  | 3  | -1.5             | 2.25                          | -4               | 16                            | 6                                |
| 3  | 8  | 0.5              | .25                           | 1                | 1                             | 0.5                              |
| 4  | 10 | 1.5              | 2.25                          | 3                | 9                             | 4.5                              |
| 10 | 28 |                  | 5.00                          |                  | 26                            | 11.0                             |

$$\bar{x} = \frac{10}{4} = 2.5 \quad \bar{y} = \frac{28}{4} = 7$$

$$s_x = \sqrt{\frac{5}{3}} = 1.2910$$

$$s_y = \sqrt{\frac{26}{3}} = 2.9439$$

$$r = \frac{\sum(X - \bar{X})(y - \bar{y})}{(n - 1)s_x s_y} = \frac{11}{(4 - 1)(1.2910)(2.9439)} = 0.9648$$

d. There is a strong correlation between the advertising expense and sales.

13-2  $H_0: \rho \leq 0, H_1: \rho > 0$ .  $H_0$  is rejected if  $t > 1.714$ .

$$t = \frac{.43 \sqrt{25 - 2}}{\sqrt{1 - (.43)^2}} = 2.284$$

$H_0$  is rejected. There is a positive correlation between the percent of the vote received and the amount spent on the campaign.

13-3 a. See the calculations in Self-Review 13-1, part (c).

$$b = \frac{rs_y}{s_x} = \frac{(0.9648)(2.9439)}{1.2910} = 2.2$$

$$a = \frac{28}{4} - 2.2 \left( \frac{10}{4} \right) = 7 - 5.5 = 1.5$$

b. The slope is 2.2. This indicates that an increase of \$1 million in advertising will result in an increase of \$2.2 million in sales. The intercept is 1.5. If there was no expenditure for advertising, sales would be \$1.5 million.

c.  $\hat{Y} = 1.5 + 2.2(3) = 8.1$

13-4  $H_0: \beta_1 \leq 0; H_1: \beta > 0$ . Reject  $H_0$  if  $t > 3.182$ .

$$t = \frac{2.2 - 0}{0.4243} = 5.1850$$

Reject  $H_0$ . The slope of the line is greater than 0.

13-5 a.

| y  | $\hat{y}$ | (y - $\hat{y}$ ) | (y - $\hat{y}$ ) <sup>2</sup> |
|----|-----------|------------------|-------------------------------|
| 7  | 5.9       | 1.1              | 1.21                          |
| 3  | 3.7       | -0.7             | .49                           |
| 8  | 8.1       | -0.1             | .01                           |
| 10 | 10.3      | -0.3             | .09                           |
|    |           |                  | 1.80                          |

$$s_{y \cdot x} = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}} = \sqrt{\frac{1.80}{4 - 2}} = .9487$$

b.  $r^2 = (.9648)^2 = .9308$

c. Ninety-three percent of the variation in sales is accounted for by advertising expense.

13-6 6.58 and 9.62, since for an x of 3 is 8.1, found by  $\hat{y} = 1.5 + 2.2(3) = 8.1$ , then  $\bar{x} = 2.5$  and  $\sum(x - \bar{x})^2 = 5$ . t from Appendix B.5 for 4 - 2 = 2 degrees of freedom at the .10 level is 2.920.

$$\hat{y} \pm t(s_{y \cdot x}) \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}}$$

$$= 8.1 \pm 2.920(0.9487) \sqrt{\frac{1}{4} + \frac{(3 - 2.5)^2}{5}}$$

$$= 8.1 \pm 2.920(0.9487)(0.5477)$$

$$= 6.58 \text{ and } 9.62 \text{ (in \$ millions)}$$

### CHAPTER 14

14-1 a. \$389,500 or 389.5 (in \$000); found by

$$2.5 + 3(40) + 4(72) - 3(10) + .2(20) + 1(5) = 3,895$$

b. The  $b_2$  of 4 shows profit will go up \$4,000 for each extra hour the restaurant is open (if none of the other variables change). The  $b_3$  of -3 implies profit will fall \$3,000 for each added mile away from the SkyWheel (if none of the other variables change).

- 14-2 a.** The total degrees of freedom ( $n - 1$ ) is 25, so the sample size is 26.  
**b.** There are 5 independent variables.  
**c.** There is only 1 dependent variable (profit).  
**d.**  $S_{Y,12345} = 1.414$ , found by  $\sqrt{2}$ . Ninety-five percent of the residuals will be between  $-2.828$  and  $2.828$ , found by  $\pm 2(1.414)$ .  
**e.**  $R^2 = .714$ , found by  $100/140$ . 71.4% of the deviation in profit is accounted for by these five variables.  
**f.**  $R_{adj}^2 = .643$ , found by

$$1 - \left[ \frac{40}{(26 - (5 + 1))} \right] \Big/ \left[ \frac{140}{(26 - 1)} \right]$$

- 14-3 a.**  $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$   
 $H_1$ : Not all of the  $\beta$ s are 0.  
 The decision rule is to reject  $H_0$  if  $F > 2.71$ . The computed value of  $F$  is 10, found by  $20/2$ . So, you reject  $H_0$ , which indicates at least one of the regression coefficients is different from zero.

Based on  $p$ -values, the decision rule is to reject the null hypothesis if the  $p$ -value is less than .05. The computed value of  $F$  is 10, found by  $20/2$ , and has a  $p$ -value of .000. Thus, we reject the null hypothesis, which indicates that at least one of the regression coefficients is different from zero.

- b.** For variable 1:  $H_0: \beta_1 = 0$  and  $H_1: \beta_1 \neq 0$   
 The decision rule is to reject  $H_0$  if  $t < -2.086$  or  $t > 2.086$ . Since 2.000 does not go beyond either of those limits, we fail to reject the null hypothesis. This regression coefficient could be zero. We can consider dropping this variable. By parallel logic, the null hypothesis is rejected for variables 3 and 4.

For variable 1, the decision rule is to reject  $H_0: \beta_1 = 0$  if the  $p$ -value is less than .05. Because the  $p$ -value is .056, we cannot reject the null hypothesis. This regression coefficient could be zero. Therefore, we can consider dropping this variable. By parallel logic, we reject the null hypothesis for variables 3 and 4.

- c.** We should consider dropping variables 1, 2, and 5. Variable 5 has the smallest absolute value of  $t$  or largest  $p$ -value. So delete it first and compute the regression equation again.

**14-4 a.**  $\hat{y} = 15.7625 + 0.4415x_1 + 3.8598x_2$   
 $\hat{y} = 15.7625 + 0.4415(30) + 3.8598(1)$   
 $= 32.87$

- b.** Female agents make \$3,860 more than male agents.

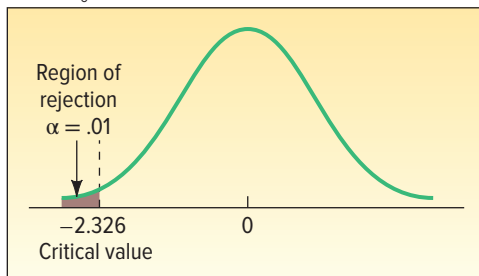
- c.**  $H_0: \beta_3 = 0$   
 $H_1: \beta_3 \neq 0$   
 $df = 17$ ; reject  $H_0$  if  $t < -2.110$  or  $t > 2.110$

$$t = \frac{3.8598 - 0}{1.4724} = 2.621$$

The  $t$ -statistic exceeds the critical value of 2.110. Also, the  $p$ -value = .0179 and is less than .05. Reject  $H_0$ . Gender should be included in the regression equation.

## CHAPTER 15

- 15-1 a.** Yes, because both  $n\pi$  and  $n(1 - \pi)$  exceed 5:  $n\pi = 200(.40) = 80$ , and  $n(1 - \pi) = 200(.60) = 120$ .  
**b.**  $H_0: \pi \geq .40$   
 $H_1: \pi < .40$   
**c.** Reject  $H_0$  if  $z < -2.326$ .



- d.**  $z = -0.87$ , found by:

$$z = \frac{.37 - .40}{\sqrt{\frac{.40(1 - .40)}{200}}} = \frac{-.03}{\sqrt{.0012}} = -0.87$$

Do not reject  $H_0$ .

- e.** The  $p$ -value is .1922, found by  $.5000 - .3078$ .

- 15-2 a.**  $H_0: \pi_o = \pi_{ch}$

$$H_1: \pi_o \neq \pi_{ch}$$

- b.** .10

- c.** Two-tailed

- d.** Reject  $H_0$  if  $z < -1.645$  or  $z > 1.645$ .

**e.**  $p_c = \frac{87 + 123}{150 + 200} = \frac{210}{350} = .60$

$$p_o = \frac{87}{150} = .58 \quad p_{ch} = \frac{123}{200} = .615$$

$$z = \frac{.58 - .615}{\sqrt{\frac{.60(.40)}{150} + \frac{.60(.40)}{200}}} = -0.66$$

- f.** Do not reject  $H_0$ .

**g.**  $p\text{-value} = 2(.5000 - .2454) = .5092$

There is no difference in the proportion of adults and children that liked the proposed flavor.

- 15-3 a.** Observed frequencies

- b.** Six (six days of the week)

- c.** 10. Total observed frequencies  $\div 6 = 60/6 = 10$ .

- d.** 5;  $k - 1 = 6 - 1 = 5$

- e.** 15.086 (from the chi-square table in Appendix B.7).

**f.**  $\chi^2 = \sum \left[ \frac{(f_o - f_e)^2}{f_e} \right] = \frac{(12 - 10)^2}{10} + \dots + \frac{(9 - 10)^2}{10} = 0.8$

- g.** Do not reject  $H_0$ .

- h.** Evidence fails to show a difference in the proportion of absences by day of the week.

- 15-4**  $H_0: P_C = .60, P_L = .30, \text{ and } P_U = .10$ .

$H_1$ : Distribution is not as above.

Reject  $H_0$  if  $\chi^2 > 5.991$ .

| Category      | $f_o$ | $f_e$ | $\frac{(f_o - f_e)^2}{f_e}$ |
|---------------|-------|-------|-----------------------------|
| Current       | 320   | 300   | 1.33                        |
| Late          | 120   | 150   | 6.00                        |
| Uncollectible | 60    | 50    | 2.00                        |
|               | 500   | 500   | 9.33                        |

Reject  $H_0$ . The accounts receivable data do not reflect the national average.

- 15-5 a.** Contingency table

- b.**  $H_0$ : There is no relationship between income and whether the person played the lottery.

$H_1$ : There is a relationship between income and whether the person played the lottery.

- c.** Reject  $H_0$  if  $\chi^2 > 5.991$ .

**d.**  $\chi^2 = \frac{(46 - 40.71)^2}{40.71} + \frac{(28 - 27.14)^2}{27.14} + \frac{(21 - 27.14)^2}{27.14}$   
 $+ \frac{(14 - 19.29)^2}{19.29} + \frac{(12 - 12.86)^2}{12.86} + \frac{(19 - 12.86)^2}{12.86}$   
 $= 6.544$

- e.** Reject  $H_0$ . There is a relationship between income level and playing the lottery.

# Glossary

**Alternate hypothesis** A statement that is accepted if the sample data provide sufficient evidence that the null hypothesis is false.

**Analysis of variance (ANOVA)** A technique used to test simultaneously whether the means of several populations are equal. It uses the  $F$  distribution as the distribution of the test statistic.

**Autocorrelation** Successive residuals in a time series are correlated.

**Bar chart** A graph that shows qualitative classes on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are proportional to the heights of the bars.

**Binomial probability distribution** A probability distribution based on a discrete random variable. Its major characteristics are: 1. An outcome on each trial of an experiment is classified into one of two mutually exclusive categories—a success or a failure. 2. The random variable is the number of successes in a fixed number of trials. 3. The probability of success is the same for each trial. 4. The trials are independent, meaning that the outcome of one trial does not affect the outcome of any other trial.

**Box plot** A graphic display that shows the general shape of a variable's distribution. It is based on five descriptive statistics: the maximum and minimum values, the first and third quartiles, and the median.

**Cause-and-effect diagram** A diagram used to illustrate the relationship between a problem and a set of the problem's possible causes.

**Central limit theorem** If all samples of a particular size are selected from any population, the sampling distribution of the sample mean is approximately a normal distribution. This approximation improves with larger samples.

**Chebyshev's theorem** For any set of observations (sample or population), the proportion of the values that lie within  $k$  standard deviations of the mean is at least  $1 - 1/k^2$ , where  $k$  is any value greater than 1.

**Classical probability** Probability based on the assumption we know the number of possible outcomes and that each of the outcomes is equally likely.

**Cluster sampling** A population is divided into clusters using naturally occurring geographic or other boundaries. Then, clusters are randomly selected and a sample is collected by randomly selecting from each cluster.

**Coefficient of determination** The proportion of the total variation in the dependent variable  $Y$  that is explained, or accounted for, by the variation in the independent variable  $X$ .

**Coefficient of multiple determination** The percent of variation in the dependent variable,  $y$ , explained by the set of independent variables,  $x_1, x_2, x_3, \dots, x_k$ .

**Collectively exhaustive** At least one of the events must occur when an experiment is conducted.

**Combination formula** A formula to count the number of possible arrangements when the order of the outcomes is not important. For example, the outcome {a, b, c} is considered the same as {c, b, a}.

**Conditional probability** The probability of a particular event occurring, given that another event has occurred.

**Confidence interval** A range of values constructed from sample data so that the population parameter is likely to occur within that range at a specified probability. The specified probability is called the *level of confidence*.

**Contingency table** A table used to classify sample observations according to two identifiable characteristics.

**Continuous random variable** A random variable that may assume an infinite number of values within a given range.

**Correlation analysis** A group of techniques to measure the relationship between two variables.

**Correlation coefficient** A measure of the strength of the linear relationship between two variables.

**Critical value** The dividing point between the region where the null hypothesis is rejected and the region where it is not rejected.

**Deciles** Values of an ordered (minimum to maximum) data set that divide the data into 10 equal parts.

**Dependent variable** The variable that is being predicted or estimated.

**Descriptive statistics** Methods of organizing, summarizing, and presenting data in an informative way.

**Discrete random variable** A random variable that can assume only certain clearly separated values.

**Dot plot** A dot plot summarizes the distribution of one variable by stacking dots at points on a number line that shows the values of the variable. A dot plot shows all values.

**Dummy variable** A variable in which there are only two possible outcomes. For analysis, one of the outcomes is coded a 1 and the other a 0.

**Empirical probability** The probability of an event happening is the fraction of the time similar events happened in the past.

**Empirical Rule** For a symmetrical, bell-shaped frequency distribution, approximately 68% of the observations lie within plus and minus one standard deviation of the mean; about 95% of the observations lie within plus and minus two standard deviations of the mean; and practically all (99.7%) lie within plus and minus three standard deviations of the mean.

**Event** A collection of one or more outcomes of an experiment.

**Experiment** A process that leads to the occurrence of one and only one of several possible results.

**Frequency distribution** A grouping of quantitative data into mutually exclusive and collectively exhaustive classes showing the number of observations in each class.

**Frequency table** A grouping of qualitative data into mutually exclusive and collectively exhaustive classes showing the number of observations in each class.

**Global test** A test used to determine if any of the set of independent variables has regression coefficients different from zero.

**Histogram** A graph in which the classes are marked on the horizontal axis and the class frequencies on the vertical axis.

The class frequencies are represented by the heights of the bars, and the bars are drawn adjacent to each other.

**Homoscedasticity** The variation around the regression equation is the same for all of the values of the independent variables.

**Hypothesis** A statement about a population parameter subject to verification.

**Hypothesis testing** A procedure based on sample evidence and probability theory to determine whether the hypothesis is a reasonable statement.

**Independence** The occurrence of one event has no effect on the probability of the occurrence of another event.

**Independent variable** A variable that provides the basis for estimation.

**Inferential statistics** The methods used to estimate a property of a population on the basis of a sample.

**Interquartile range** The absolute numerical difference between the first and third quartiles. Fifty percent of a distribution's values occur in this range.

**Interval level of measurement** For data recorded at the interval level of measurement, the interval or the distance between values is meaningful. The interval level of measurement is based on a scale with a known unit of measurement.

**Joint probability** A probability that measures the likelihood two or more events will happen concurrently.

**Law of large numbers** Over a large number of trials, the empirical probability of an event will approach its true probability.

**Least squares principle** A mathematical procedure that uses the data to position a line with the objective of minimizing the sum of the squares of the vertical distances between the actual  $y$  values and the predicted values of  $y$ .

**Level of significance** The probability of rejecting the null hypothesis when it is true.

**Measure of dispersion** A value that shows the spread of a data set. The range, variance, and standard deviation are measures of dispersion.

**Measure of location** A single value that is typical of the data. It pinpoints the center of a distribution. The arithmetic mean, weighted mean, median, mode, and geometric mean are measures of location.

**Median** The midpoint of the values after they have been ordered from the minimum to the maximum values.

**Mode** The value of the observation that appears most frequently.

**Multiplication formula** If there are  $m$  ways of doing one thing and  $n$  ways of doing another thing, there are  $m \times n$  ways of doing both.

**Mutually exclusive** The occurrence of one event means that none of the other events can occur at the same time.

**Nominal level of measurement** Data recorded at the nominal level of measurement are represented as labels or names. They have no order. They can only be classified and counted.

**Null hypothesis** A statement about the value of a population parameter developed for the purpose of testing numerical evidence.

**Ordinal level of measurement** Data recorded at the ordinal level of measurement are based on a relative ranking or rating of items based on a defined attribute or qualitative variable. Variables based on this level of measurement are only ranked or counted.

**Outcome** A particular result of an experiment.

**Outlier** A data point that is unusually far from the others. An accepted rule is to classify an observation as an outlier if it is 1.5 times the interquartile range above the third quartile or below the first quartile.

**$p$ -value** The probability of observing a sample value as extreme as, or more extreme than, the value observed, given that the null hypothesis is true.

**Parameter** A characteristic of a population.

**Percentiles** Values of an ordered (minimum to maximum) data set that divide the data into 100 intervals.

**Permutation** Any arrangement of  $r$  objects selected from a single group of  $n$  possible objects.

**Permutation formula** A formula to count the number of possible arrangements when the order of the outcomes is important. For example, the outcome  $\{a, b, c\}$  is considered different from  $\{c, b, a\}$ .

**Pie chart** A chart that shows the proportion or percentage that each class represents of the total number of frequencies.

**Point estimate** The statistic, computed from sample information, that estimates a population parameter.

**Poisson experiment** An experiment where the random variable is the number of times an outcome is observed in a clearly defined interval. Examples of an interval include time, distance, area, or volume. In addition, the experiment requires that the probability of an outcome is proportional to the length of the interval, and the outcomes in each interval are independent.

**Poisson probability distribution** A mathematical function used to calculate the probabilities of a Poisson experiment, where the random variable is the number of times an outcome occurs in a clearly defined interval, the probability of an outcome is proportional to the length of the interval, and the intervals are independent.

**Population** The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest.

**Probability** A value between 0 and 1, inclusive, describing the relative possibility (chance or likelihood) an event will occur.

**Probability distribution** A listing of all the outcomes of an experiment and the probability associated with each outcome.

**Proportion** The fraction, ratio, or percent indicating the part of the sample or the population having a particular trait of interest.

**Qualitative variables** A nominal-scale variable coded to assume only one nonnumeric outcome or category. For example, a person is considered either employed or unemployed.

**Quartiles** Values of an ordered (minimum to maximum) data set that divide the data into four intervals.

**Random variable** A variable measured or observed as the result of an experiment. By chance, the variable can have different values.

**Random variation** The sum of the squared differences between each observation and its treatment mean.

**Range** A measure of dispersion found by subtracting the minimum value from the maximum value.

**Ratio level of measurement** Data recorded at the ratio level of measurement are based on a scale with a known unit of measurement and a meaningful interpretation of zero on the scale.

**Regression equation** An equation that expresses the linear relationship between two variables.

**Residual** The difference between the actual value of the dependent variable and the estimated value of the dependent variable, that is,  $y - \hat{y}$ .

**Sample** A portion, or part, of the population of interest.

**Sampling distribution of the sample mean** A probability distribution of all possible sample means of a given sample size.

**Sampling error** The difference between a sample statistic and its corresponding population parameter.

**Scatter diagram** Graphical technique used to show the relationship between two variables measured with interval or ratio scales.

**Simple random sample** A sample selected so that each item or person in the population has the same chance of being included.

**Special rule of addition** A rule used to find the probabilities of events made up of  $A$  or  $B$  when the events are mutually exclusive.

**Special rule of multiplication** A rule used to find the probability of the joint occurrence of independent events.

**Standard error of estimate** A measure of the dispersion, or scatter, of the observed values around the line of regression for a given value of  $x$ .

**Statistic** A characteristic of a sample.

**Statistics** The science of collecting, organizing, presenting analyzing, and interpreting data to assist in making more effective decisions.

**Stepwise regression** A step-by-step method to determine a regression equation that begins with a single independent variable and adds or deletes independent variables one by one. Only independent variables with nonzero regression coefficients are included in the regression equation.

**Stratified random sample** A population is divided into subgroups, called strata, and a sample is randomly selected from each stratum.

**Subjective concept of probability** The likelihood (probability) of a particular event happening that is assigned by an individual based on whatever information is available.

**Systematic random sampling** A random starting point is selected, and then every  $k$ th member of the population is selected.

**Test statistic** A value, determined from sample information, used to decide whether to reject the null hypothesis.

**Total variation** The sum of the squared differences between each observation and the overall mean.

**Treatment variation** The sum of the squared differences between each treatment mean and the grand or overall mean. Each squared difference is multiplied by the number of observations in the treatment.

**Type I error** Rejecting the null hypothesis,  $H_0$ , when it is true.

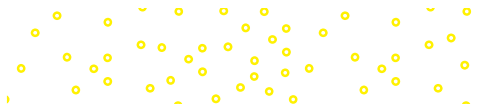
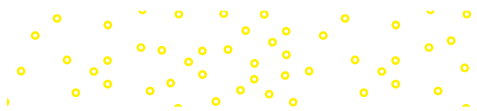
**Type II error** Not rejecting the null hypothesis when it is false.

**Variance** The arithmetic mean of the squared deviations from the mean.

**Variance inflation factor** A test used to detect correlation among independent variables.

**z value** The signed distance between a selected value, designated  $x$ , and the mean,  $\mu$ , divided by the standard deviation,  $\sigma$ . Also called *z score*.

# Index

- 
- Able Moving and Storage Inc., 470
- Addition rules  
  general, 129–131  
  special, 126–127
- Adjusted coefficient of determination, 428
- Alpha ( $\alpha$ ), 278
- Alternate hypothesis  
  explanation of, 277  
  one-tailed test and, 282, 286  
  two-tailed test and, 283–286
- American Association of Retired Persons (AARP), 470
- American Hospital Administrators Association (AHAA), 486
- American Restaurant Association, 242
- Analysis of variance (ANOVA)  
  applications for, 341–342  
  assumptions and, 340–342  
  explanation of, 335  
  *F* distribution and, 335–339  
  inferences about pairs of treatment means and, 350–352
- ANOVA tables  
  coefficient of determination and, 394, 427  
  multiple regression and, 425, 426, 430, 431  
  use of, 389, 394
- ANOVA test  
  explanation of, 342–348  
  use of, 351–352
- Arithmetic mean. *See* Mean
- Asymptotic distributions, 190, 430. *See also* Normal probability distributions
- Attributes, measurement of, 9
- Autocorrelation, 441–442
- AutoNation, 20
- Averages, 54
- Backward elimination method, 447
- Bank of New England, 418
- Bar charts, 22
- Bell-shaped distributions, 79, 190. *See also* Normal probability distributions
- Best Buy, 1
- Best-subset regression, 435, 447
- Beta ( $\beta$ ), 278, 380, 388
- Bias, 215
- Bill of Mortality, 5
- Bimodal distributions, 100
- Binomial probability distributions  
  application of, 191  
  cumulative, 171–172  
  explanation of, 164–165  
  graphs of, 169  
  mean of, 167  
  method to compute, 165–167  
  normal approximation to, 471  
  Poisson distributions and, 176–177  
  table of, 513–517
- Binomial probability formula, 165–167
- Binomial probability tables, 167–168
- Bivariate data, 89, 104
- BMW, 20
- Box plots, 96–99
- Burger King, 260
- Bush, George W., 133
- Business analytics, statistical knowledge and, 12–13
- Cargill, 3
- Central limit theorem, 225–231
- Challenger* space shuttle, 366
- Charts. *See also* Graphic presentations  
  bar, 22, 24, 25  
  control, 23  
  pie, 22–25
- Chebyshev, P. L., 78
- Chebyshev's theorem, 78–79
- Chi-square distribution  
  characteristics of, 483  
  contingency table analysis and, 490–493  
  critical values of, 525
- Chi-square statistic, background of, 483
- Chi-square test  
  contingency table analysis and, 490–493  
  limitations of, 487–488  
  unequal expected frequencies and, 486–487
- Churchill Downs, 53
- Class frequencies  
  explanation of, 29  
  number of, 28  
  as relative class frequencies, 21
- Classical probability, 121–122
- Class intervals  
  equal, 30  
  explanation of, 28
- Class limits, 28–29
- Class midpoint, 30
- Cluster sampling, 216–217
- 

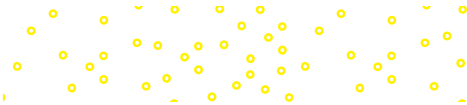
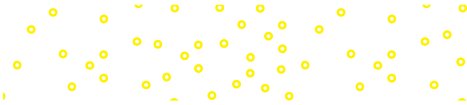
- Coefficient of determination
  - adjusted, 428
  - from ANOVA table, 394
  - explanation of, 392–394
  - formula for, 427
  - relationship to correlation coefficient and standard error of estimate, 392–394
- Coefficient of multiple determination, 427–428
- Collectively exhaustive events, 122
- Combination formula, 145–146
- Combinations, 144
- Comparable studies, 264
- Complement rule, 128–129
- Conditional probability, 134, 137, 138–139
- Confidence intervals
  - computer simulation and, 249–251
  - construction of, 397–400
  - for difference in treatment means, 350–352
  - explanation of, 243, 244, 275, 397
  - for mean, 397
  - 90%, 245, 247
  - 95%, 245, 246, 248, 262, 351
  - for population mean, 244–248
  - for population proportion, 260–263
  - population standard deviation and, 244–248
  - sample size and, 263–269
  - use of, 254–259
- Contingency tables
  - analysis of, 490–493
  - explanation of, 106–107, 135–136
- Continuous probability distributions
  - area under normal curve and, 191, 192, 196–201
  - explanation of, 185
  - normal, 189–191
  - standard normal, 192–201
  - uniform, 185–188
- Continuous random variables
  - explanation of, 160
  - uniform probability distributions and, 185–186
- Continuous variables, 7
- Control charts, 197
- Correlation
  - perfect, 369
  - simple linear, 419
  - spurious, 373
- Correlation analysis
  - explanation of, 366–367
  - scatter diagrams and, 367–368, 371, 373
- Correlation coefficient
  - characteristics of, 370–372
  - explanation of, 369–370
  - formula for, 372
  - interpretation of, 372–374
  - method to determine value of, 370–372
  - Pearson product-moment, 369
  - relationship to coefficient of determination and standard error of estimate, 392–394
  - square of, 394
  - testing significance of, 376–378
  - t* test for, 377
- Correlation matrix, 449
- Counting principles
  - combination formula as, 145–146
  - explanation of, 142
  - multiplication formula as, 142–143
  - permutation formula as, 143–144
- Critical value
  - chi-square distribution, 525
  - explanation of, 280
  - F* distribution and, 337, 523–524
  - for one-tailed and two-tailed tests, 287
- Cumulative binomial probability distributions, 171–172
- Cumulative frequency distributions, 39–41
- Cumulative frequency polygons, 39–41
- Data
  - bivariate, 89, 104
  - classified by levels of measurement, 7–11
  - collection and generation of, 2–3
  - interval-level, 9–10
  - nominal-level, 7–8
  - ordinal-level, 8–9
  - ratio-level, 10–11
  - raw, 56
  - univariate, 104
- Deciles
  - calculation of, 94
  - explanation of, 92
- Decision rule, 279–280, 284–285, 292, 377
- Decision trees. See Tree diagrams
- Degrees of freedom (*df*), 425
- Dell, 3
- Del Monte Foods, 252
- Dependent events, 133
- Dependent samples
  - independent samples vs., 321–323
  - two-sample tests of, 318–321

- Dependent variables  
  explanation of, 368  
  regression analysis and, 368, 419, 430  
  (See *also* Regression analysis)  
  stock price and, 380
- Descriptive statistics, 4, 20, 118
- Deviation. See Standard deviation
- Discrete probability distributions  
  binomial probability distributions as,  
    164–169, 191  
  cumulative binomial probability distributions as,  
    171–172  
  mean of, 160  
  Poisson, 173–177  
  variance and standard deviation of, 160–162
- Discrete random variables, 159
- Discrete variables, 7
- Dispersion. See Measures of dispersion
- Distribution-free tests. See Hypothesis tests
- Distributions. See Continuous probability distributions; Frequency distributions; Probability distributions
- Dole Pineapple Inc., 274
- Dot plots, 89–90
- Dummy variable, 442
- Empirical probability, 122–123
- Empirical Rule, 79–80, 193–195
- Enron, 12
- Environmental Protection Agency (EPA), 243
- Error. See Sampling error; Standard error; Type I error; Type II error
- Ethical issues, 12, 81
- Events  
  collectively exhaustive, 122  
  dependent, 133  
  explanation of, 120  
  inclusive, 130  
  independent, 132–133  
  joint, 129–130  
  mutually exclusive, 20, 21, 121–122, 126, 127
- Excel (Microsoft)  
  ANOVA, 348, 431  
  area under normal curve, 197  
  combinations, 146  
  confidence intervals, 257–258  
  correlation coefficient, 372, 374, 377  
  frequency distributions, 31–32  
  frequency tables, 21, 23  
  histograms, 34  
  mean, median, and mode, 66–67, 71, 77  
  multiple regression, 421–422, 433, 449, 451  
  probability distribution, 168, 172, 203  
  qualitative independent variables, 442  
  quartiles, 94–95  
  random sampling, 213–214  
  regression analysis, 384–385, 397  
  scatter diagrams, 105, 106  
  skewness, 101  
  standard deviation, 77  
  statistical summary, 13  
  *t* tests, 315, 321  
  use of, 12–13, 20, 21, 23, 31
- Expected frequencies, 479–483, 486–487, 492
- Expected value, 160
- Experiments, 119–120
- F* distributions  
  characteristics of, 335–336  
  comparing two population variances,  
    335–339  
  global test and, 430  
  table of, 523–524  
  as test statistic, 340  
  use of, 340
- The Federalist*, 30
- Fisher, R. A., 213
- Fisher, Ronald, 335
- Fitbit, 1
- Forbes*, 3
- Ford Motor Company, 20, 490
- Forward selection method, 447. See *also* Stepwise regression
- Frequencies  
  class, 21, 28, 29, 31  
  equal expected, 479–483  
  expected, 479–483, 486–487, 492  
  relative class, 21  
  unequal expected, 486–487
- Frequency distributions  
  construction of, 28–31  
  cumulative, 39–41  
  cumulative frequency polygons as, 39–41  
  explanation of, 27, 89, 118  
  frequency polygons as, 36–37  
  graphic presentation of, 33–37  
  histograms for, 33–35, 64  
  negatively skewed, 65, 100–101  
  positively skewed, 64, 65  
  relative, 31–32, 122, 123  
  skewed, 64, 100–101



- Frequency polygons  
 cumulative, 39–41  
 explanation of, 36–37  
 shapes of, 100
- Frequency tables  
 construction of, 21  
 explanation of, 20  
 pie and bar charts and, 24  
 relative, 21
- Frito-Lay, 4
- F*-statistic  
 null hypothesis and, 431  
 as test statistic, 336, 340, 430
- Gates, William, 3
- General Foods Corporation, 281
- General Motors, 20, 470, 474
- General multiple regression equation, 419–420.  
*See also* Multiple regression equation
- General rule of addition, 129–131
- General rule of multiplication, 133–134
- Geometric mean, as measure of location, 54
- Gibbs Baby Food Company, 305
- Global test, 430, 431
- Goodness-of-fit tests  
 equal expected frequencies and, 479–483  
 limitations of, 487–488  
 unequal expected frequencies and,  
 486–487  
 use of, 479
- Google, 2, 334
- Gosset, William, 252
- Gould, Stephen Jay, 101
- Graphic presentations. *See also* Charts  
 box plots as, 96–99  
 contingency tables as, 106–107, 135–138  
 cumulative frequency polygons as, 41  
 dot plots as, 89–90  
 of frequency distributions, 33–37  
 frequency polygons as, 36–37  
 histograms as, 33–35, 64, 453  
 of qualitative data, 22–26  
 residual plots as, 436–437  
 scatter diagrams as, 105, 106, 367–368, 371, 373,  
 436–437, 452  
 tree diagrams as, 138–140  
 Venn diagrams as, 127–131
- Graunt, John, 5
- Group 1 Automotive Inc., 20
- Guinness Brewery, 252
- Gwynn, Tony, 76
- Hamilton, Alexander, 30
- Histograms  
 symmetrical distribution and, 64  
 use of, 33–35, 64, 453
- Homoscedasticity, 438
- Hypotheses  
 alternate, 277, 282, 283–284  
 explanation of, 275–276  
 null, 276–287
- Hypothesis tests. *See also* Analysis of variance  
 (ANOVA); Nonparametric methods  
 correlation coefficient and, 376–378  
 of equal expected frequencies, 479–483  
 equal population variances and, 335–339  
 explanation of, 275–276, 470  
 Minitab and, 296–297  
 one-tailed, 281–282, 286–287  
 for population mean with known standard deviation,  
 283–286  
 for population mean with unknown standard  
 deviation, 290–294  
*p*-values in, 287–288, 321, 377  
 significance of slope and, 388  
 six-step procedure for, 276–281, 338,  
 344–345  
*t* distribution and, 290, 340–341, 389  
 two-sample, 306–321 (*See also* Two-sample  
 hypothesis tests)  
 two-tailed, 281–286  
 Type I error and, 278, 279, 281  
 Type II error and, 278  
 of unequal expected frequencies,  
 486–487
- Hyundai, 20
- Inclusive events, 130
- Independent events, 132–133
- Independent samples. *See also* Two-sample  
 hypothesis tests  
 dependent samples vs., 321–323  
 two-sample hypothesis tests and, 306–311
- Independent variables  
 explanation of, 368  
 multicollinearity and, 439–441  
 qualitative, 442–444  
 regression analysis and, 368, 419, 430, 442–444  
 (*See also* Regression analysis)  
 stock market and, 380
- Inferential statistics, 5
- Interquartile range, 97
- Interval-level data, 9–10

- Jay, John, 30  
Joint events, 129–130  
Joint probability, 129–130, 137, 138
- Kennedy, John F., 94  
Kentucky Derby, 53  
Kia, 20  
Koch Industries, 3  
Kroger, 2
- Landon, Alfred, 215  
LASIK, 281  
Law of large numbers, 122–123  
Least squares method  
    explanation of, 380–383, 420  
    regression line and, 384–385, 399  
Level of confidence, 264  
Level of significance, 277–279, 284, 291  
Linear regression  
    assumptions underlying, 396–397  
    drawing regression line and, 383–386  
    least squares principle and, 380–384  
    multiple, 429–435  
    prediction intervals and, 396–400  
    testing significance of slope and,  
        388–390  
    use of, 419  
Linear regression equation, 382  
*Literary Digest* poll (1936), 215  
Location. *See* Measures of location  
Lockheed Martin, 366  
Lotteries, 487
- Madison, James, 30  
Madoff, Bernie, 12  
Margin of error, 262, 264  
Martin Marietta, 366  
McGivern Jewelers, 88  
Mean  
    applications for, 190  
    of binomial probability distribution, 167  
    of discrete probability distribution, 160  
    distribution shape and, 100  
    Empirical Rule and, 193–195  
    issues in use of, 58  
    as measure of location, 54  
    of normal distribution, 190, 191  
    of Poisson distribution, 174, 177  
    properties of, 57–58  
    relative position and, 64–65  
    sample, 56–57  
    skewness and, 100–101  
    weighted, 67  
Mean square, 347  
Mean square error (MSE), 347, 350  
Mean square for treatments (MST), 347  
Measurement levels  
    interval, 9–10  
    nominal, 7–8  
    ordinal, 8–9  
    ratio, 10–11  
    summary and examples of, 11  
Measures of dispersion  
    purpose of, 54  
    range as, 69–70  
    reasons to study, 68–69  
    standard deviation (*See* Standard deviation)  
    variance as, 70–72  
Measures of location  
    formula for, 93  
    mean as, 54–58  
    median as, 59–61  
    mode as, 61–62  
    purpose of, 54, 68  
    relative positions of mean, median, and mode and,  
        64–65  
    software example, 66–67  
    types of, 54  
Measures of position  
    formula for, 93  
    purpose of, 92  
    quartiles, deciles, and percentiles and,  
        92–95  
Median  
    distribution shape and, 100  
    explanation of, 59–60  
    as measure of location, 54  
    properties of, 61  
    relative position and, 64–65  
    skewness and, 100–101  
MegaStat. *See also* Excel (Microsoft)  
    best-subset regression and, 447  
    chi-square test, 488, 489, 493  
    quartiles, 94  
    two-sample test of proportions, 477  
    use of, 12, 31  
Merrill Lynch, 19  
Method of least squares. *See* Least squares method  
Microsoft Corporation, 3  
Microsoft Excel. *See* Excel (Microsoft)

- 
- Minitab  
box plots, 98  
confidence intervals, 257  
correlation coefficient, 372, 377  
dot plots, 90  
one-sample hypothesis tests, 296–297  
one-way ANOVA, 351  
prediction intervals, 399  
quartiles, 94  
random samples, 213  
relationship between variables, 402  
skewness, 101–103  
stepwise regression, 445–447  
use of, 12, 31
- Mode  
disadvantages of using, 62  
explanation of, 61–62  
as measure of location, 54  
relative position and, 64–65
- Model, of relationship, 429
- Monroe, Marilyn, 54
- Morton Thiokol, 366
- Multicollinearity, 439–441
- Multiple linear regression, 429–435
- Multiple regression analysis  
autocorrelation and, 441–442  
background on, 419–420  
distribution of residuals and, 439  
evaluating assumptions of, 436–442  
example of, 420–422  
homoscedasticity and, 438  
independent observations and, 441–442  
linear relationships and, 436–437  
multicollinearity and, 439–441  
qualitative independent variables and, 442–444  
review of, 448–453  
stepwise regression and, 435, 445–447  
uses for, 419, 442
- Multiple regression equation  
adjusted coefficient of determination and, 428  
ANOVA table and, 425, 426  
coefficient of multiple determination and, 427–428  
evaluation of, 425–428  
example of, 420–422  
explanation of, 419–420, 453  
multiple standard error of estimate and, 426–427
- Multiple regression model, 430–432
- Multiple standard error of estimate, 426–427
- Multiplication formula, 142–143
- Multiplication rules  
general, 133–134  
special, 132–133
- Mutually exclusive events  
explanation of, 20, 21, 121–122  
special rule of addition and, 126, 127
- NASDAQ, 20
- National Collegiate Athletic Association (NCAA), 137
- Negatively skewed distributions, 65, 100
- New York Stock Exchange, 20
- Nightingale, Florence, 36
- Nike, 210
- 90% confidence intervals, 245, 247
- 95% confidence intervals, 245, 246, 248, 262, 351
- Nixon, Richard, 94
- Nominal-scale variables, 7–8, 470
- Nonnumeric variables, 21
- Nonparametric methods. *See also* Hypothesis tests  
background on, 470  
chi-square limitations and, 487–488  
contingency table analysis and, 490–493  
goodness-of-fit tests and, 479–483  
hypothesis test of population proportion and, 388–390  
hypothesis test of unexpected frequencies, 486–487  
two-sample tests about proportion and, 474–477
- Normal approximation to binomial distribution, 471
- Normal curve  
continuous probability distributions and area under, 191, 192, 196–198, 202–203  
finding area under, 196–201  
table of area under, 519
- Normal probability distributions  
area under curve and, 192, 196–198, 202–203  
characteristics of, 190–191  
combining two areas and, 199–200  
converted to standard, 192  
family of, 190  
formula for, 189  
means and, 190, 191  
residuals and, 439  
standard, 192–201  
standard deviation and, 79, 191
- Normal probability plot, 439
- Null hypothesis  
decision rule and, 279–280  
explanation of, 276–277  
hypothesis test result and, 280–281  
level of significance and, 277–279
- 

- multiple regression and, 430
- one-tailed and two-tailed tests and, 281–286
- Numeric data. *See* Quantitative variables
- One-sample hypothesis tests for population mean
  - with known standard deviation, 283–286
  - with unknown standard deviation, 290–294
- One-tailed test
  - example of, 286–287, 389
  - explanation of, 281–282, 286–287
- One-way ANOVA, 348, 351
- Ordinal-level data, 8–9
- Outcomes, 120, 127
- Outliers, 99
- $p$ -values, 287–288, 321, 377, 431, 432, 444, 450
- Paired samples, 318
- Paired  $t$  test, 319
- Parameter, population, 55, 219
- Pearson, Karl, 100, 369, 483
- Pearson product-moment correlation
  - coefficient, 369
- Pearson's coefficient of skewness, 100–102
- Pearson's  $r$ , 369
- Penske Auto Group, 20
- Percentiles, 92–94
- Perfect correlation, 369
- Permutation formula, 143–144
- Permutations, 143
- Pie charts
  - explanation of, 22–23
  - frequency tables and, 24
  - uses for, 25
- Pilot studies, 264
- Point estimate
  - characteristics of, 173
  - explanation of, 243, 257
  - Poisson experiment, 173–174
  - for population mean, 243–244
- Poisson probability distributions
  - application of, 173–177
  - binomial probability and, 176–177
  - characteristics of, 174
  - explanation of, 174
  - formula for, 174
  - mean of, 174, 177
  - table of, 518
  - variance of, 174
- Ponzi scheme, 12
- Pooled proportion, 475–477
- Pooled variance, 313
- Population
  - explanation of, 5
  - parameter of, 55, 219
- Population mean
  - compared with unknown population standard deviations, 312–321
  - confidence intervals for, 244–248
  - explanation of, 55–56
  - hypothesis tests for, 283–287, 290–294
  - point estimate for, 243–244
  - sample size to estimate, 264–265
  - two-tailed test for, 283–286
  - unbiased estimator of, 220
- Population parameter, 219
- Population proportion
  - confidence interval for, 260–263
  - hypothesis tests for, 470–473
  - sample size to estimate, 265–266
- Population standard deviation
  - explanation of, 75, 244–248
  - known, 244–248, 283–287
  - sample size and, 264
  - unknown, 252–259, 290–294
- Population variance, 73–74
- Position. *See* Measures of position
- Positively skewed distributions, 64, 65, 227–229, 336, 430
- Practically significant, 287
- Prediction intervals, 397–400
- Probability
  - approaches to, 121–122
  - classical, 121–122
  - conditional, 134, 137, 138
  - counting principles and, 142–146
  - empirical, 122–123
  - explanation of, 119–120
  - joint, 129–130, 137, 138
  - subjective, 124
- Probability distributions. *See also* Continuous probability distributions; Discrete probability distributions; Uniform probability distributions
  - application of, 160, 211
  - binomial, 164–169
  - characteristics of, 156
  - cumulative binomial, 171–172
  - explanation of, 156
  - $F$  distributions (*See F* distributions)
  - generation of, 156–158
  - mean of, 160
  - Poisson, 173–177
  - random variables and, 158–160
  - variance and standard deviation of, 160–162

- Probability rules  
 complement rule of, 128–129  
 general rule of addition as, 129–131  
 general rule of multiplication as, 133–134  
 special rule of addition as, 126–127  
 special rule of multiplication as, 132–133
- Proportions  
 confidence intervals for, 260–263  
 pooled, 475–477  
 population, 260–263, 265–266,  
 470–473  
 sample, 470  
 two-sample tests of, 474–477
- Pseudo-random numbers, 213
- Qualitative variables  
 explanation of, 6, 7, 21  
 in graphic form, 22–26  
 in multiple regression, 442–444  
 ordinal-level data and, 8–9
- Quality control. *See* Control charts
- Quantitative variables  
 continuous, 7  
 discrete, 7  
 explanation of, 6, 7, 21  
 measures of location to describe,  
 54–67
- Quartiles  
 box plots and, 96–99  
 calculation of, 94–95  
 explanation of, 92, 93
- RAND Corporation, 213
- Random numbers  
 in lotteries, 487  
 pseudo-, 213
- Random numbers table, 212, 520
- Random samples. *See also* Samples/sampling  
 simple, 212–214  
 statistical software to create, 249–251  
 stratified, 215–216  
 systematic, 215
- Random variables  
 continuous, 160, 185–186  
 discrete, 159  
 explanation of, 158–159
- Random variation, 342–343
- Range, 69–70
- Range-based approach, 264
- Ratio-level data, 10–11
- Raw data, 55, 56
- Regression analysis. *See also* Linear regression;  
 Multiple regression analysis  
 drawing regression line and, 383–386  
 explanation of, 366, 380  
 least squares method and, 380–384  
 transformation and, 400–403
- Regression coefficients  
 evaluation of, 432–435  
 explanation of, 380, 422  
 testing individual, 432, 450
- Regression equation. *See also* Multiple regression  
 equation  
 ability to predict and, 391–395  
 explanation of, 380  
 general form of, 382  
 hypothesis tests to analyze,  
 388–390  
 interval estimates of prediction and,  
 396–400  
 method to determine, 383, 388  
 multiple, 432–434  
 test of hypothesis to analyze,  
 388–390
- Regression line  
 explanation of, 419  
 least squares, 384, 399  
 method to draw, 383–384  
 slope of, 382
- Relative class frequencies, 21, 31
- Relative frequency distributions, 31–32,  
 122, 123
- Relative frequency tables  
 discrete random variables and, 159  
 frequency tables converted to, 21  
 pie and bar charts and, 24
- Residual plots, 436–437
- Residuals  
 calculation of, 385  
 distribution of, 439  
 variation in, 438
- Risk, regression analysis to quantify, 380
- Roosevelt, Franklin D., 215
- R*-square, 392
- Rules of probability. *See* Probability rules
- Sample mean. *See also* Sampling distribution of  
 sample mean  
 central limit theorem and, 225–231  
 explanation of, 56–57  
 formula for, 56
- Sample proportion, formula to compute, 470

- Sample size
  - confidence intervals and, 263–264
  - to estimate population mean, 264–265
  - to estimate population proportion, 265–266
- Samples/sampling
  - central limit theorem and, 225–231
  - cluster, 216–217
  - dependent, 321–323
  - determining size of, 243
  - explanation of, 5, 211
  - independent, 306–311, 321–323
  - paired, 318
  - point estimate for population mean and, 243–244
  - reasons for, 211–212, 243
  - simple random, 212–214
  - stratified random, 215–216
  - systematic random, 215
  - use of, 5
- Sample standard deviation, 77
- Sample statistic, 219
- Sample variance, 76–77
- Sampling distribution of sample mean
  - central limit theorem and, 225–231
  - explanation of, 221–223
  - population standard deviation and, 245
  - use of, 221–223, 232–233
- Sampling error
  - example of, 250–251
  - explanation of, 219–220, 227
- Scatter diagrams
  - correlation analysis and, 367–368, 371, 373
  - multiple regression and, 436–437, 452
  - use of, 105, 106
- Simple random samples, 212–214
- Skewed distributions
  - explanation of, 64, 65
  - positively, 64, 65, 227–229
- Skewness
  - calculation of, 100–103
  - explanation of, 100
  - Pearson's coefficient of, 100–103
  - software coefficient of, 101
- Slope
  - of regression line, 382
  - testing significance of, 388–390
- Software, statistical, 12–13. *See also* Excel (Microsoft); MegaStat; Minitab
- Software coefficient of skewness, 101
- Southern Technical Institute, 260
- Special rule of addition, 126–127
- Special rule of multiplication, 132–133
- Spurious correlation, 373
- Standard deviation
  - Chebyshev's theorem and, 78–79
  - of discrete probability distribution, 160–162
  - Empirical Rule and, 79–80, 193–195
  - explanation of, 244–245
  - interpretation and use of, 78–80
  - normal probability distributions and, 79, 190, 191
  - population, 75, 244–248, 252–259, 264, 283–287
  - sample, 77
  - of uniform distribution, 186
- Standard error, 245
- Standard error of estimate
  - calculation of, 391–392
  - explanation of, 391
  - formula for, 391
  - multiple, 426–427
  - prediction and, 395
  - relationship to coefficients of correlation and determination, 392–394
- Standard error of mean, 230
- Standardizing, 101
- Standard mean, 244
- Standard normal probability distribution
  - applications of, 193
  - areas under normal curve and, 196–201
  - Empirical Rule and, 193–195
  - explanation of, 191, 192
  - normal probability distribution converted into, 192
- Standard normal table, 246
- Standard normal value, 192
- Standard & Poor's 500 Index, 380
- State Farm Mutual Automobile Insurance, 3
- Statistic
  - explanation of, 56, 57
  - sample, 219
- Statistical inference
  - applications for, 6, 118
  - explanation of, 5, 118, 211, 275
  - multiple regression analysis and, 429–435
  - pairs of treatment means and, 350–352
  - sampling and, 215

- Statistically significant, 287
- Statistics
- descriptive, 4, 20, 118
  - ethics and, 12
  - explanation of, 3–4
  - history of, 2–3
  - inferential, 5–6, 118
  - misleading, 12
  - reasons to study, 2–3, 12–13
- Stepwise regression, 435, 445–447
- Stock market, 380
- Strata, 215
- Stratified random samples, 215–216
- Student's *t* distribution, 253, 258, 521–522
- Subjective probability, 124
- Sum of squares error (SSE), 394, 425
- Sum of squares total (SS total), 394
- Symbols, pronunciation and meaning of, 82, 110
- Symmetric distributions, 64, 100, 190. *See also* Normal probability distributions
- Systematic random samples, 215
- t* distribution
- characteristics of, 252–253
  - confidence interval for population mean and, 253–257
  - development of, 252
  - hypothesis testing and, 290, 340–341, 389
  - Student's, 253, 258, 521–522
- t* tests
- for correlation coefficient, 377
  - Excel procedure for, 315
  - paired, 319
- Table of random numbers, 212–213
- Target, 2
- Television viewership, 469
- Test statistic
- for comparing two variances, 335–339
  - explanation of, 279, 284, 291
- Tippett, L., 213
- Total variation, 342
- Travelair.com, 365
- Treatment means, inferences about pairs of, 350–352
- Treatments, 342–343
- Treatment variation, 342–343
- Tree diagrams, 138–140
- Tukey, John W., 94
- Two-sample hypothesis tests
- dependent samples and, 318–321
  - independent samples and, 306–311
  - of means - known  $\sigma$ , 308
  - two-sample pooled test and, 312–316
- Two-sample pooled test, 312–316
- Two-sample tests
- of means, 313
  - of proportions, 474–477
- Two-tailed test
- critical value of *F* for, 337
  - example of, 283–286
  - explanation of, 281–282
- Tyco, 12
- Type I error
- example of, 281
  - explanation of, 278, 279
  - statistical software and, 351
- Type II error, 278, 279
- Unbiased sampling methods, 215
- Unequal expected frequencies, 486–487
- Uniform probability distributions
- equation for, 186
  - examples of, 186–188
  - explanation of, 185–186
  - standard deviation of, 186
- Univariate data, 104
- University of Michigan Institute for Social Research, 422
- U.S. Postal Service, 69
- Van Tuyl Group, 20
- Variables
- continuous, 7
  - continuous random, 160, 185–186
  - dependent, 368 (*See also* Dependent variables)
  - dummy, 442
  - independent, 368 (*See also* Independent variables)
  - nominal-scale, 7–8, 470
  - nonnumeric, 21
  - qualitative, 6–8, 22–26
  - quantitative, 6, 7, 21, 54–67
  - random, 158–160
  - relationship between two, 104–106, 366, 367 (*See also* Correlation analysis)
  - types of, 6–7
- Variance. *See also* Analysis of variance (ANOVA)
- of binomial probability distribution, 167
  - of discrete probability distribution, 160–162
  - of distribution of differences in means, 308

- explanation of, 70–72
- of Poisson distribution, 174
- pooled, 313
- population, 73–74, 335–339
- sample, 76–77
- Variance inflation factor (VIF), 440, 441
- Variation
  - random, 342–343
  - in residuals, 438
  - total, 342
  - treatment, 342–343
- Venn, J., 127
- Venn diagrams, 127–131
- Volvo, 20
- Walmart, 2
- Weighted mean, 54, 67
- Yates, F., 213
- Y-intercept, 382
- z distribution, use of, 252, 253, 335
- z values (z scores), 192, 198, 232, 233, 245, 258–259, 266







**KEY FORMULAS** Lind, Marchal, and Wathen • *Basic Statistics for Business and Economics*, 9th edition

**CHAPTER 3**

- Population mean

$$\mu = \frac{\sum X}{N} \quad (3-1)$$

- Sample mean, raw data

$$\bar{x} = \frac{\sum X}{n} \quad (3-2)$$

- Weighted mean

$$\bar{x}_w = \frac{w_1X_1 + w_2X_2 + \dots + w_nX_n}{w_1 + w_2 + \dots + w_n} \quad (3-3)$$

- Range

$$\text{Range} = \text{Maximum value} - \text{Minimum value} \quad (3-4)$$

- Population variance

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N} \quad (3-5)$$

- Population standard deviation

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}} \quad (3-6)$$

- Sample variance

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} \quad (3-7)$$

- Sample standard deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \quad (3-8)$$

**CHAPTER 4**

- Location of a percentile

$$L_p = (n + 1) \frac{P}{100} \quad (4-1)$$

- Pearson's coefficient of skewness

$$sk = \frac{3(\bar{x} - \text{Median})}{s} \quad (4-2)$$

- Software coefficient of skewness

$$sk = \frac{n}{(n - 1)(n - 2)} \left[ \sum \left( \frac{x - \bar{x}}{s} \right)^3 \right] \quad (4-3)$$

**CHAPTER 5**

- Special rule of addition

$$P(A \text{ or } B) = P(A) + P(B) \quad (5-2)$$

- Complement rule

$$P(A) = 1 - P(-A) \quad (5-3)$$

- General rule of addition

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \quad (5-4)$$

- Special rule of multiplication

$$P(A \text{ and } B) = P(A)P(B) \quad (5-5)$$

- General rule of multiplication

$$P(A \text{ and } B) = P(A)P(B|A) \quad (5-6)$$

- Multiplication formula

$$\text{Total number of arrangements} = (m)(n) \quad (5-7)$$

- Permutation formula

$${}_n P_r = \frac{n!}{(n - r)!} \quad (5-8)$$

- Combination formula

$${}_n C_r = \frac{n!}{r!(n - r)!} \quad (5-9)$$

**CHAPTER 6**

- Mean of a probability distribution

$$\mu = \sum [xP(x)] \quad (6-1)$$

- Variance of a probability distribution

$$\sigma^2 = \sum [(x - \mu)^2 P(x)] \quad (6-2)$$

- Binomial probability distribution

$$P(x) = {}_n C_x \pi^x (1 - \pi)^{n - x} \quad (6-3)$$

- Mean of a binomial distribution

$$\mu = n\pi \quad (6-4)$$

- Variance of a binomial distribution

$$\sigma^2 = n\pi(1 - \pi) \quad (6-5)$$

- Poisson probability distribution

$$P(x) = \frac{\mu^x e^{-\mu}}{x!} \quad (6-6)$$

- Mean of a Poisson distribution

$$\mu = n\pi \quad (6-7)$$

## CHAPTER 7

- Mean of a uniform distribution

$$\mu = \frac{a + b}{2} \quad (7-1)$$

- Standard deviation of a uniform distribution

$$\sigma = \sqrt{\frac{(b - a)^2}{12}} \quad (7-2)$$

- Uniform probability distribution

$$P(x) = \frac{1}{b - a} \quad (7-3)$$

if  $a \leq x \leq b$  and 0 elsewhere

- Normal probability distribution

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]} \quad (7-4)$$

- Standard normal value

$$z = \frac{x - \mu}{\sigma} \quad (7-5)$$

## CHAPTER 8

- Standard error of the mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (8-1)$$

- z-value,  $\mu$  and  $\sigma$  known

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad (8-2)$$

## CHAPTER 9

- Confidence interval for  $\mu$ , with  $\sigma$  known

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}} \quad (9-1)$$

- Confidence interval for  $\mu$ ,  $\sigma$  unknown

$$\bar{x} \pm t \frac{s}{\sqrt{n}} \quad (9-2)$$

- Sample proportion

$$p = \frac{x}{n} \quad (9-3)$$

- Confidence interval for proportion

$$p \pm z \sqrt{\frac{p(1-p)}{n}} \quad (9-4)$$

- Sample size for estimating mean

$$n = \left(\frac{z\sigma}{E}\right)^2 \quad (9-5)$$

- Sample size for a proportion

$$n = \pi(1 - \pi) \left(\frac{z}{E}\right)^2 \quad (9-6)$$

## CHAPTER 10

- Testing a mean,  $\sigma$  known

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad (10-1)$$

- Testing a mean,  $\sigma$  unknown

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad (10-2)$$

## CHAPTER 11

- Variance of the distribution of difference in means

$$\sigma_{\bar{x}_1 - \bar{x}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad (11-1)$$

- Two-sample test of means, known  $\sigma$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (11-2)$$

- Pooled variance

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (11-3)$$

- Two-sample test of means, unknown but equal  $\sigma$ 's

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (11-4)$$

- Paired  $t$  test

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} \quad (11-5)$$

## CHAPTER 12

- Test for comparing two variances

$$F = \frac{s_1^2}{s_2^2} \quad (12-1)$$

- Sum of squares, total

$$SS \text{ total} = \Sigma(x - \bar{x}_G)^2 \quad (12-2)$$

- Sum of squares, error

$$SSE = \Sigma(x - \bar{x}_c)^2 \quad (12-3)$$

- Sum of squares, treatments

$$SST = SS \text{ total} - SSE \quad (12-4)$$

- Confidence interval for differences in treatment means

$$(\bar{x}_1 - \bar{x}_2) \pm t \sqrt{\text{MSE} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (12-5)$$

## CHAPTER 13

- Correlation coefficient

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{(n - 1) s_x s_y} \quad (13-1)$$

- Test for significant correlation

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (13-2)$$

- Linear regression equation

$$\hat{y} = a + bx \quad (13-3)$$

- Slope of the regression line

$$b = r \frac{s_y}{s_x} \quad (13-4)$$

- Intercept of the regression line

$$a = \bar{y} - b\bar{x} \quad (13-5)$$

- Test for a zero slope

$$t = \frac{b - 0}{s_b} \quad (13-6)$$

- Standard error of estimate

$$s_{y \cdot x} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}} \quad (13-7)$$

- Coefficient of determination

$$r^2 = \frac{SSR}{SS \text{ Total}} = 1 - \frac{SSE}{SS \text{ Total}} \quad (13-8)$$

- Standard error of estimate

$$s_{y \cdot x} = \sqrt{\frac{SSE}{n - 2}} \quad (13-9)$$

- Confidence interval

$$\hat{y} \pm t_{s_{y \cdot x}} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma(x - \bar{x})^2}} \quad (13-11)$$

- Prediction interval

$$\hat{y} \pm t_{s_{y \cdot x}} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma(x - \bar{x})^2}} \quad (13-12)$$

## CHAPTER 14

- Multiple regression equation

$$\hat{y} = a + b_1x_1 + b_2x_2 + \dots + b_kx_k \quad (14-1)$$

- Multiple standard error of estimate

$$S_{y \cdot 123 \dots k} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - (k + 1)}} = \sqrt{\frac{SSE}{n - (k + 1)}} \quad (14-2)$$

- Coefficient of multiple determination

$$R^2 = \frac{SSR}{SS \text{ total}} \quad (14-3)$$

- Adjusted coefficient of determination

$$R_{adj}^2 = 1 - \frac{\frac{SSE}{n - (k + 1)}}{\frac{SS \text{ total}}{n - 1}} \quad (14-4)$$

- Global test of hypothesis

$$F = \frac{SSR/k}{SSE/[n - (k + 1)]} \quad (14-5)$$

- Testing for a particular regression coefficient

$$t = \frac{b_l - 0}{s_{b_l}} \quad (14-6)$$

- Variance inflation factor

$$VIF = \frac{1}{1 - R_j^2} \quad (14-7)$$

## CHAPTER 15

- Test of hypothesis, one proportion

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} \quad (15-1)$$

- Two-sample test of proportions

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_c(1 - p_c)}{n_1} + \frac{p_c(1 - p_c)}{n_2}}} \quad (15-2)$$

- Pooled proportion

$$p_c = \frac{x_1 + x_2}{n_1 + n_2} \quad (15-3)$$

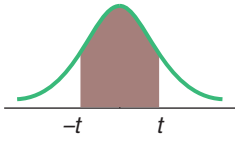
- Chi-square test statistic

$$\chi^2 = \sum \left[ \frac{(f_o - f_e)^2}{f_e} \right] \quad (15-4)$$

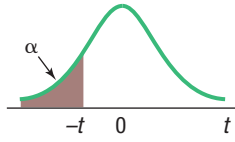
- Expected frequency

$$f_e = \frac{(\text{Row total})(\text{Column total})}{\text{Grand total}} \quad (15-5)$$

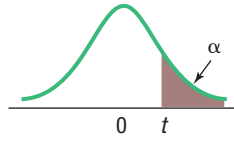
# Student's *t* Distribution



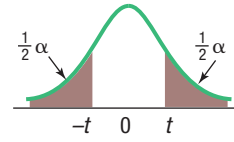
Confidence interval



Left-tailed test



Right-tailed test



Two-tailed test

(continued)

| df<br>(degrees<br>of<br>freedom)                    | Confidence Intervals, <i>c</i>                      |       |        |        |        |         |
|---|---|-------|--------|--------|--------|---------|
|   | 80%   | 90%   | 95%    | 98%    | 99%    | 99.9%   |
|   | Level of Significance for One-Tailed Test, $\alpha$ |       |        |        |        |         |
|   | 0.10  | 0.05  | 0.025  | 0.01   | 0.005  | 0.0005  |
| Level of Significance for Two-Tailed Test, $\alpha$ |   |       |        |        |        |         |
| 0.20  | 0.10  | 0.05  | 0.02   | 0.01   | 0.001  |         |
| 1   | 3.078   | 6.314 | 12.706 | 31.821 | 63.657 | 636.619 |
| 2   | 1.886   | 2.920 | 4.303  | 6.965  | 9.925  | 31.599  |
| 3   | 1.638   | 2.353 | 3.182  | 4.541  | 5.841  | 12.924  |
| 4   | 1.533   | 2.132 | 2.776  | 3.747  | 4.604  | 8.610   |
| 5   | 1.476   | 2.015 | 2.571  | 3.365  | 4.032  | 6.869   |
| 6   | 1.440   | 1.943 | 2.447  | 3.143  | 3.707  | 5.959   |
| 7   | 1.415   | 1.895 | 2.365  | 2.998  | 3.499  | 5.408   |
| 8   | 1.397   | 1.860 | 2.306  | 2.896  | 3.355  | 5.041   |
| 9   | 1.383   | 1.833 | 2.262  | 2.821  | 3.250  | 4.781   |
| 10  | 1.372   | 1.812 | 2.228  | 2.764  | 3.169  | 4.587   |
| 11  | 1.363   | 1.796 | 2.201  | 2.718  | 3.106  | 4.437   |
| 12  | 1.356   | 1.782 | 2.179  | 2.681  | 3.055  | 4.318   |
| 13  | 1.350   | 1.771 | 2.160  | 2.650  | 3.012  | 4.221   |
| 14  | 1.345   | 1.761 | 2.145  | 2.624  | 2.977  | 4.140   |
| 15  | 1.341   | 1.753 | 2.131  | 2.602  | 2.947  | 4.073   |
| 16  | 1.337   | 1.746 | 2.120  | 2.583  | 2.921  | 4.015   |
| 17  | 1.333   | 1.740 | 2.110  | 2.567  | 2.898  | 3.965   |
| 18  | 1.330   | 1.734 | 2.101  | 2.552  | 2.878  | 3.922   |
| 19  | 1.328   | 1.729 | 2.093  | 2.539  | 2.861  | 3.883   |
| 20  | 1.325   | 1.725 | 2.086  | 2.528  | 2.845  | 3.850   |
| 21  | 1.323   | 1.721 | 2.080  | 2.518  | 2.831  | 3.819   |
| 22  | 1.321   | 1.717 | 2.074  | 2.508  | 2.819  | 3.792   |
| 23  | 1.319   | 1.714 | 2.069  | 2.500  | 2.807  | 3.768   |
| 24  | 1.318   | 1.711 | 2.064  | 2.492  | 2.797  | 3.745   |
| 25  | 1.316   | 1.708 | 2.060  | 2.485  | 2.787  | 3.725   |
| 26  | 1.315   | 1.706 | 2.056  | 2.479  | 2.779  | 3.707   |
| 27  | 1.314   | 1.703 | 2.052  | 2.473  | 2.771  | 3.690   |
| 28  | 1.313   | 1.701 | 2.048  | 2.467  | 2.763  | 3.674   |
| 29  | 1.311   | 1.699 | 2.045  | 2.462  | 2.756  | 3.659   |
| 30  | 1.310   | 1.697 | 2.042  | 2.457  | 2.750  | 3.646   |
| 31  | 1.309   | 1.696 | 2.040  | 2.453  | 2.744  | 3.633   |
| 32  | 1.309   | 1.694 | 2.037  | 2.449  | 2.738  | 3.622   |
| 33  | 1.308   | 1.692 | 2.035  | 2.445  | 2.733  | 3.611   |
| 34  | 1.307   | 1.691 | 2.032  | 2.441  | 2.728  | 3.601   |
| 35  | 1.306   | 1.690 | 2.030  | 2.438  | 2.724  | 3.591   |

(continued-top right)

| df<br>(degrees<br>of<br>freedom)                    | Confidence Intervals, <i>c</i>                      |       |       |       |       |        |
|---|---|-------|-------|-------|-------|--------|
|   | 80%   | 90%   | 95%   | 98%   | 99%   | 99.9%  |
|   | Level of Significance for One-Tailed Test, $\alpha$ |       |       |       |       |        |
|   | 0.10  | 0.05  | 0.025 | 0.01  | 0.005 | 0.0005 |
| Level of Significance for Two-Tailed Test, $\alpha$ |   |       |       |       |       |        |
| 0.20  | 0.10  | 0.05  | 0.02  | 0.01  | 0.001 |        |
| 36  | 1.306   | 1.688 | 2.028 | 2.434 | 2.719 | 3.582  |
| 37  | 1.305   | 1.687 | 2.026 | 2.431 | 2.715 | 3.574  |
| 38  | 1.304   | 1.686 | 2.024 | 2.429 | 2.712 | 3.566  |
| 39  | 1.304   | 1.685 | 2.023 | 2.426 | 2.708 | 3.558  |
| 40  | 1.303   | 1.684 | 2.021 | 2.423 | 2.704 | 3.551  |
| 41  | 1.303   | 1.683 | 2.020 | 2.421 | 2.701 | 3.544  |
| 42  | 1.302   | 1.682 | 2.018 | 2.418 | 2.698 | 3.538  |
| 43  | 1.302   | 1.681 | 2.017 | 2.416 | 2.695 | 3.532  |
| 44  | 1.301   | 1.680 | 2.015 | 2.414 | 2.692 | 3.526  |
| 45  | 1.301   | 1.679 | 2.014 | 2.412 | 2.690 | 3.520  |
| 46  | 1.300   | 1.679 | 2.013 | 2.410 | 2.687 | 3.515  |
| 47  | 1.300   | 1.678 | 2.012 | 2.408 | 2.685 | 3.510  |
| 48  | 1.299   | 1.677 | 2.011 | 2.407 | 2.682 | 3.505  |
| 49  | 1.299   | 1.677 | 2.010 | 2.405 | 2.680 | 3.500  |
| 50  | 1.299   | 1.676 | 2.009 | 2.403 | 2.678 | 3.496  |
| 51  | 1.298   | 1.675 | 2.008 | 2.402 | 2.676 | 3.492  |
| 52  | 1.298   | 1.675 | 2.007 | 2.400 | 2.674 | 3.488  |
| 53  | 1.298   | 1.674 | 2.006 | 2.399 | 2.672 | 3.484  |
| 54  | 1.297   | 1.674 | 2.005 | 2.397 | 2.670 | 3.480  |
| 55  | 1.297   | 1.673 | 2.004 | 2.396 | 2.668 | 3.476  |
| 56  | 1.297   | 1.673 | 2.003 | 2.395 | 2.667 | 3.473  |
| 57  | 1.297   | 1.672 | 2.002 | 2.394 | 2.665 | 3.470  |
| 58  | 1.296   | 1.672 | 2.002 | 2.392 | 2.663 | 3.466  |
| 59  | 1.296   | 1.671 | 2.001 | 2.391 | 2.662 | 3.463  |
| 60  | 1.296   | 1.671 | 2.000 | 2.390 | 2.660 | 3.460  |
| 61  | 1.296   | 1.670 | 2.000 | 2.389 | 2.659 | 3.457  |
| 62  | 1.295   | 1.670 | 1.999 | 2.388 | 2.657 | 3.454  |
| 63  | 1.295   | 1.669 | 1.998 | 2.387 | 2.656 | 3.452  |
| 64  | 1.295   | 1.669 | 1.998 | 2.386 | 2.655 | 3.449  |
| 65  | 1.295   | 1.669 | 1.997 | 2.385 | 2.654 | 3.447  |
| 66  | 1.295   | 1.668 | 1.997 | 2.384 | 2.652 | 3.444  |
| 67  | 1.294   | 1.668 | 1.996 | 2.383 | 2.651 | 3.442  |
| 68  | 1.294   | 1.668 | 1.995 | 2.382 | 2.650 | 3.439  |
| 69  | 1.294   | 1.667 | 1.995 | 2.382 | 2.649 | 3.437  |
| 70  | 1.294   | 1.667 | 1.994 | 2.381 | 2.648 | 3.435  |

(continued)

# Student's *t* Distribution (concluded)

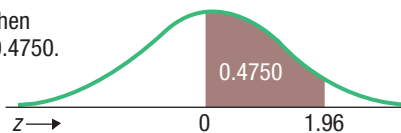
(continued)

| <i>df</i><br>(degrees<br>of<br>freedom) | Confidence Intervals, <i>c</i>                      |       |       |       |       |        |
|---|---|-------|-------|-------|-------|--------|
|   | 80%   | 90%   | 95%   | 98%   | 99%   | 99.9%  |
|   | Level of Significance for One-Tailed Test, $\alpha$ |       |       |       |       |        |
|   | 0.10  | 0.05  | 0.025 | 0.01  | 0.005 | 0.0005 |
|   | Level of Significance for Two-Tailed Test, $\alpha$ |       |       |       |       |        |
|   | 0.20  | 0.10  | 0.05  | 0.02  | 0.01  | 0.001  |
| 71                                      | 1.294   | 1.667 | 1.994 | 2.380 | 2.647 | 3.433  |
| 72                                      | 1.293   | 1.666 | 1.993 | 2.379 | 2.646 | 3.431  |
| 73                                      | 1.293   | 1.666 | 1.993 | 2.379 | 2.645 | 3.429  |
| 74                                      | 1.293   | 1.666 | 1.993 | 2.378 | 2.644 | 3.427  |
| 75                                      | 1.293   | 1.665 | 1.992 | 2.377 | 2.643 | 3.425  |
| 76                                      | 1.293   | 1.665 | 1.992 | 2.376 | 2.642 | 3.423  |
| 77                                      | 1.293   | 1.665 | 1.991 | 2.376 | 2.641 | 3.421  |
| 78                                      | 1.292   | 1.665 | 1.991 | 2.375 | 2.640 | 3.420  |
| 79                                      | 1.292   | 1.664 | 1.990 | 2.374 | 2.640 | 3.418  |
| 80                                      | 1.292   | 1.664 | 1.990 | 2.374 | 2.639 | 3.416  |
| 81                                      | 1.292   | 1.664 | 1.990 | 2.373 | 2.638 | 3.415  |
| 82                                      | 1.292   | 1.664 | 1.989 | 2.373 | 2.637 | 3.413  |
| 83                                      | 1.292   | 1.663 | 1.989 | 2.372 | 2.636 | 3.412  |
| 84                                      | 1.292   | 1.663 | 1.989 | 2.372 | 2.636 | 3.410  |
| 85                                      | 1.292   | 1.663 | 1.988 | 2.371 | 2.635 | 3.409  |
| 86                                      | 1.291   | 1.663 | 1.988 | 2.370 | 2.634 | 3.407  |
| 87                                      | 1.291   | 1.663 | 1.988 | 2.370 | 2.634 | 3.406  |
| 88                                      | 1.291   | 1.662 | 1.987 | 2.369 | 2.633 | 3.405  |
| 89                                      | 1.291   | 1.662 | 1.987 | 2.369 | 2.632 | 3.403  |
| 90                                      | 1.291   | 1.662 | 1.987 | 2.368 | 2.632 | 3.402  |
| 91                                      | 1.291   | 1.662 | 1.986 | 2.368 | 2.631 | 3.401  |
| 92                                      | 1.291   | 1.662 | 1.986 | 2.368 | 2.630 | 3.399  |
| 93                                      | 1.291   | 1.661 | 1.986 | 2.367 | 2.630 | 3.398  |
| 94                                      | 1.291   | 1.661 | 1.986 | 2.367 | 2.629 | 3.397  |
| 95                                      | 1.291   | 1.661 | 1.985 | 2.366 | 2.629 | 3.396  |
| 96                                      | 1.290   | 1.661 | 1.985 | 2.366 | 2.628 | 3.395  |
| 97                                      | 1.290   | 1.661 | 1.985 | 2.365 | 2.627 | 3.394  |
| 98                                      | 1.290   | 1.661 | 1.984 | 2.365 | 2.627 | 3.393  |
| 99                                      | 1.290   | 1.660 | 1.984 | 2.365 | 2.626 | 3.392  |
| 100                                     | 1.290   | 1.660 | 1.984 | 2.364 | 2.626 | 3.390  |
| 120                                     | 1.289   | 1.658 | 1.980 | 2.358 | 2.617 | 3.373  |
| 140                                     | 1.288   | 1.656 | 1.977 | 2.353 | 2.611 | 3.361  |
| 160                                     | 1.287   | 1.654 | 1.975 | 2.350 | 2.607 | 3.352  |
| 180                                     | 1.286   | 1.653 | 1.973 | 2.347 | 2.603 | 3.345  |
| 200                                     | 1.286   | 1.653 | 1.972 | 2.345 | 2.601 | 3.340  |
| $\infty$                                | 1.282   | 1.645 | 1.960 | 2.326 | 2.576 | 3.291  |



## Areas under the Normal Curve

Example:  
If  $z = 1.96$ , then  
 $P(0 \text{ to } z) = 0.4750$ .



| <b>z</b> | <b>0.00</b> | <b>0.01</b> | <b>0.02</b> | <b>0.03</b> | <b>0.04</b> | <b>0.05</b> | <b>0.06</b> | <b>0.07</b> | <b>0.08</b> | <b>0.09</b> |
|----------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 0.0      | 0.0000      | 0.0040      | 0.0080      | 0.0120      | 0.0160      | 0.0199      | 0.0239      | 0.0279      | 0.0319      | 0.0359      |
| 0.1      | 0.0398      | 0.0438      | 0.0478      | 0.0517      | 0.0557      | 0.0596      | 0.0636      | 0.0675      | 0.0714      | 0.0753      |
| 0.2      | 0.0793      | 0.0832      | 0.0871      | 0.0910      | 0.0948      | 0.0987      | 0.1026      | 0.1064      | 0.1103      | 0.1141      |
| 0.3      | 0.1179      | 0.1217      | 0.1255      | 0.1293      | 0.1331      | 0.1368      | 0.1406      | 0.1443      | 0.1480      | 0.1517      |
| 0.4      | 0.1554      | 0.1591      | 0.1628      | 0.1664      | 0.1700      | 0.1736      | 0.1772      | 0.1808      | 0.1844      | 0.1879      |
| 0.5      | 0.1915      | 0.1950      | 0.1985      | 0.2019      | 0.2054      | 0.2088      | 0.2123      | 0.2157      | 0.2190      | 0.2224      |
| 0.6      | 0.2257      | 0.2291      | 0.2324      | 0.2357      | 0.2389      | 0.2422      | 0.2454      | 0.2486      | 0.2517      | 0.2549      |
| 0.7      | 0.2580      | 0.2611      | 0.2642      | 0.2673      | 0.2704      | 0.2734      | 0.2764      | 0.2794      | 0.2823      | 0.2852      |
| 0.8      | 0.2881      | 0.2910      | 0.2939      | 0.2967      | 0.2995      | 0.3023      | 0.3051      | 0.3078      | 0.3106      | 0.3133      |
| 0.9      | 0.3159      | 0.3186      | 0.3212      | 0.3238      | 0.3264      | 0.3289      | 0.3315      | 0.3340      | 0.3365      | 0.3389      |
| 1.0      | 0.3413      | 0.3438      | 0.3461      | 0.3485      | 0.3508      | 0.3531      | 0.3554      | 0.3577      | 0.3599      | 0.3621      |
| 1.1      | 0.3643      | 0.3665      | 0.3686      | 0.3708      | 0.3729      | 0.3749      | 0.3770      | 0.3790      | 0.3810      | 0.3830      |
| 1.2      | 0.3849      | 0.3869      | 0.3888      | 0.3907      | 0.3925      | 0.3944      | 0.3962      | 0.3980      | 0.3997      | 0.4015      |
| 1.3      | 0.4032      | 0.4049      | 0.4066      | 0.4082      | 0.4099      | 0.4115      | 0.4131      | 0.4147      | 0.4162      | 0.4177      |
| 1.4      | 0.4192      | 0.4207      | 0.4222      | 0.4236      | 0.4251      | 0.4265      | 0.4279      | 0.4292      | 0.4306      | 0.4319      |
| 1.5      | 0.4332      | 0.4345      | 0.4357      | 0.4370      | 0.4382      | 0.4394      | 0.4406      | 0.4418      | 0.4429      | 0.4441      |
| 1.6      | 0.4452      | 0.4463      | 0.4474      | 0.4484      | 0.4495      | 0.4505      | 0.4515      | 0.4525      | 0.4535      | 0.4545      |
| 1.7      | 0.4554      | 0.4564      | 0.4573      | 0.4582      | 0.4591      | 0.4599      | 0.4608      | 0.4616      | 0.4625      | 0.4633      |
| 1.8      | 0.4641      | 0.4649      | 0.4656      | 0.4664      | 0.4671      | 0.4678      | 0.4686      | 0.4693      | 0.4699      | 0.4706      |
| 1.9      | 0.4713      | 0.4719      | 0.4726      | 0.4732      | 0.4738      | 0.4744      | 0.4750      | 0.4756      | 0.4761      | 0.4767      |
| 2.0      | 0.4772      | 0.4778      | 0.4783      | 0.4788      | 0.4793      | 0.4798      | 0.4803      | 0.4808      | 0.4812      | 0.4817      |
| 2.1      | 0.4821      | 0.4826      | 0.4830      | 0.4834      | 0.4838      | 0.4842      | 0.4846      | 0.4850      | 0.4854      | 0.4857      |
| 2.2      | 0.4861      | 0.4864      | 0.4868      | 0.4871      | 0.4875      | 0.4878      | 0.4881      | 0.4884      | 0.4887      | 0.4890      |
| 2.3      | 0.4893      | 0.4896      | 0.4898      | 0.4901      | 0.4904      | 0.4906      | 0.4909      | 0.4911      | 0.4913      | 0.4916      |
| 2.4      | 0.4918      | 0.4920      | 0.4922      | 0.4925      | 0.4927      | 0.4929      | 0.4931      | 0.4932      | 0.4934      | 0.4936      |
| 2.5      | 0.4938      | 0.4940      | 0.4941      | 0.4943      | 0.4945      | 0.4946      | 0.4948      | 0.4949      | 0.4951      | 0.4952      |
| 2.6      | 0.4953      | 0.4955      | 0.4956      | 0.4957      | 0.4959      | 0.4960      | 0.4961      | 0.4962      | 0.4963      | 0.4964      |
| 2.7      | 0.4965      | 0.4966      | 0.4967      | 0.4968      | 0.4969      | 0.4970      | 0.4971      | 0.4972      | 0.4973      | 0.4974      |
| 2.8      | 0.4974      | 0.4975      | 0.4976      | 0.4977      | 0.4977      | 0.4978      | 0.4979      | 0.4979      | 0.4980      | 0.4981      |
| 2.9      | 0.4981      | 0.4982      | 0.4982      | 0.4983      | 0.4984      | 0.4984      | 0.4985      | 0.4985      | 0.4986      | 0.4986      |
| 3.0      | 0.4987      | 0.4987      | 0.4987      | 0.4988      | 0.4988      | 0.4989      | 0.4989      | 0.4989      | 0.4990      | 0.4990      |