STATISTICS FOR BUSINESS

Perumal Mariappan

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Statistics for Business

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Statistics for Business

Dr Perumal Mariappan, Ph.D. (Mathematics)
Ph.D. (Management)
Head
PG Department of Actuarial Science
Bishop Heber College
Trichy – 620 017

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper

International Standard Book Number-13: 978-1-138-33617-9 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

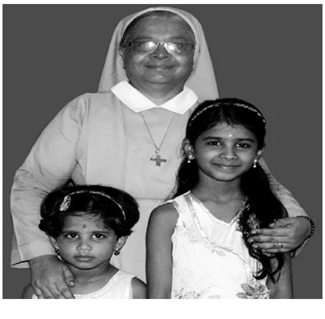
Names: Mariappan, Perumal, author.
Title: Statistics for business / Perumal Mariappan.
Description: Boca Raton, FL : CRC Press, 2019.
Identifiers: LCCN 2019010834| ISBN 9781138336179 (hardback : alk. paper) |
ISBN 9780429443244 (ebook)
Subjects: LCSH: Commercial statistics.
Classification: LCC HF1017 .M345 2019 | DDC 519.5--dc23
LC record available at https://lccn.loc.gov/2019010834

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

The author dedicates this book to Dr D. Paul Dhayabaran, Principal, Bishop Heber College, Trichy, for his wonderful dedicated service in establishing Bishop Heber College as the best higher education institution in the whole of India.

Also the author dedicates this text to Sister Siriya Pushpam, Correspondent and Former Headmistress, St. Joseph's Anglo Indian Girl's H.Sc., School, Trichy for dedicating her life to the development of the society and to his lovable daughters Kumari S. B. Bhavana Sree and Kumari S. B. Bharani Shri.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contents

Foreword xv
Preface xvii
Acknowledgements xix
Author xxi

1. Introduction to Statistics and Its Business Applications 1
1.1 Introduction 1
1.2 Is Statistics a Science? 2
1.3 Application of Statistics in Business 2
1.3.1 The Phases of the Statistical Decision-Making Process 3
1.3.1.1 Study Design Phase 3
1.3.1.2 Data Collection 4
1.3.1.3 Data Analysis 4
1.3.1.4 Action on Results 4
1.4 Responsibility of the Decision Maker 4
1.5 Functions and Limitations of Statistics 5
1.5.1 Functions of Statistics 5
1.5.2 Limitations of Statistics 5
1.6 Distrust of Statistics 6
1.7 Nature of Statistical Law 7
1.7.1 Law of Statistical Regularity 7
1.7.2 Law of Inertia of Large Numbers 7
Exercise 1 7

2. Data Structures, Data Sources, and Data Collection 9
2.1 Introduction 9
2.2 Data Structures 9
2.2.1 Univariate Data 9
2.2.2 Bivariate Data 10
2.2.3 Multivariate Data 10
2.3 Data Sources 11
2.3.1 Primary Sources 11
2.3.2 Secondary Sources 11
2.3.3 Internal Source 11
2.3.4 External Source 11
2.4 Data Collection Inquiries 12
2.4.1 Survey Design 13
2.4.1.1 Questionnaire Design 14
2.4.2 Pilot Survey of the Questionnaire 14
2.4.3 Editing Primary Data 15
2.4.4 Possible Errors in Secondary Data 15
2.4.5 Census and Sampling Methods 16
Exercise 2 19

3. Data Presentation 21
3.1 Introduction 21
3.2 Classification of Data 21
3.2.1 Types of Classification 21
3.3 Data Presentation 22
3.3.1 Textual Form 22
3.3.2 Tabular Form 22
3.4 Types of Variables and Data 23
3.5 Levels of Measurement 24
3.5.1 Nominal Scale 24
3.5.2 Ordinal Scale 24
3.5.3 Interval Scale 24
3.5.4 Ratio Scale 25
3.6 Frequency 25
3.6.1 Frequency Distributions 25
3.7 Types of Class Interval 26
3.8 Tally Mark 28
3.9 Construction of a Discrete Frequency Distribution 28
3.10 Construction of a Continuous Frequency Distribution 29
3.11 Cumulative and Relative Frequencies 29
3.12 Diagrammatic Representation of Data 30
3.12.1 Advantages and Disadvantages of Diagrammatic Representation 31
3.12.2 Types of Diagrams 31
3.12.2.1 Bar Diagram 31
3.12.2.2 Pie Diagram 34
3.12.2.3 Histogram, Frequency Polygon, and Frequency Curve 36
3.12.2.4 Frequency Polygon 37
3.12.2.5 Frequency Curve 37
3.12.2.6 Ogive Curve 38
3.12.2.7 Line Diagram 38
Exercise 3 39

4. Measures of Central Tendency (MCT) 43
4.1 Introduction 43
4.2 MCT 43
4.2.1 Properties of Best Average 43
4.3 Arithmetic Mean 44
4.3.1 Discrete Data 44
4.3.2 Discrete Data with Frequency 44
4.3.3 Continuous Data with Frequency 45
4.4 Mathematical Properties of Arithmetic Mean 46
4.5 Median 50
4.5.1 Discrete Data 51
4.5.2 Discrete Data with Frequency 51
4.5.3 Continuous Data with Frequency 51
4.5.3.1 Relative Advantages 52
4.5.3.2 Relative Disadvantages 52
4.5.3.3 Property of Median 56
4.5.4 Graphical Method to Find the Median 56

4.6 Quartiles, Deciles and Percentiles ... 58
4.7 Mode ... 61
4.7.1 Discrete Data ... 61
4.7.2 Discrete Data with Frequency ... 61
4.7.3 Continuous Data with Frequency ... 61
4.7.4 A Graphical Method to Evaluate the Mode ... 68
4.8 Comparison of Mean, Median, and Mode ... 69
4.9 Weighted Arithmetic Mean ... 70
4.9.1 Advantages of the Weighted Mean ... 71
4.10 Geometric Mean ... 71
4.11 Harmonic Mean ... 72
Exercise 4 ... 73

5. Dispersion ... 79
5.1 Introduction ... 79
5.2 Range ... 79
5.3 Quartile Deviation (QD) ... 80
5.4 Coefficient of Quartile Deviation ... 80
5.5 Mean Deviation ... 82
5.6 Standard Deviation (SD) ... 84
5.7 Relative Measures of Dispersion ... 89
Exercise 5 ... 93

6. Skewness, Moments, and Kurtosis ... 97
6.1 Introduction ... 97
6.2 Dispersion and Skewness ... 97
6.3 Moments ... 100
6.4 Kurtosis ... 102
Exercise 6 ... 105

7. Correlation and Regression Analysis ... 107
7.1 Introduction ... 107
7.2 Correlation ... 107
7.2.1 Simple Correlation or Correlation ... 107
7.2.2 Rank Correlation ... 108
7.2.3 Group Correlation ... 108
7.2.4 Assumptions for Karl Pearson's Coefficient of Correlation ... 108
7.2.5 Limitations ... 108
7.2.6 Properties ... 108
7.2.7 Scatter Diagram ... 108
7.3 Karl Pearson Coefficient of Correlation ... 109
7.4 Coefficient of Correlation of a Grouped Data ... 115
7.5 Probable Error of the Coefficient of Correlation ... 118
7.6 Rank Correlation ... 120
7.7 Regression Equations ... 125
Exercise 7 ... 134

8. Probability ... 139
8.1 Introduction ... 139
8.2 Definitions for Certain Key Terms ... 139
8.2.1 Experiment ... 139
8.2.2 Sample Space ... 139
8.2.3 Event ... 139
8.2.4 Equally Likely Events ... 140
8.2.5 Mutually Exclusive Events ... 140
8.2.6 Outcome ... 140
8.3 Meaning of Probability ... 140
8.3.1 The Classical Approach ... 140
8.3.2 The Relative Frequency Approach ... 141
8.3.3 Notation ... 141
8.3.4 Addition Rules for Probability ... 143
8.3.5 Multiplication Rule on Probability When Events Are Independent ... 146
8.3.6 Compound Probability or Conditional Probability ... 147
8.4 Bayes' Theorem ... 150
Exercise 8 ... 156

9. Random Variables and Expectation ... 161
9.1 Introduction ... 161
9.2 Random Variable ... 161
9.2.1 Discrete Random Variable ... 161
9.2.2 Continuous Random Variable ... 161
9.3 Probability Distribution ... 162
9.3.1 Discrete Probability Distribution ... 162
9.3.2 Characteristics of a Discrete Distribution ... 163
9.3.3 Probability Function ... 163
9.4 Mathematical Expectation ... 165
Exercise 9 ... 168

10. Discrete Probability Distribution: Binomial and Poisson Distributions ... 169
10.1 Introduction ... 169
10.2 Binomial Distribution ... 169
10.2.1 Characteristics of a Bernoulli Process ... 169
10.2.2 Definition of Binomial Distribution ... 169
10.2.3 Conditions of Binomial Distribution ... 170
10.2.4 Properties of Binomial Distributions ... 170
10.2.5 Mean of Binomial Distribution ... 170
10.2.6 Variance of Binomial Distribution ... 171
10.3 Poisson Distribution ... 173
10.3.1 Definition of Poisson Distribution ... 174
10.3.2 Properties of Poisson Distribution ... 174
10.3.3 Mean of the Poisson Distribution ... 174
10.3.4 Variance of the Poisson Distribution ... 174
Exercise 10 ... 178

11. Continuous Probability Distribution: Normal Distribution 179
11.1 Introduction........ 179
11.2 Definition of Normal Distribution........ 179
11.3 Standard Normal Distribution 181
11.4 Properties of Normal Distribution........ 182
Exercise 11 187

12. Theory of Sampling 189
12.1 Introduction........ 189
12.2 Why Sample? 189
12.3 How to Choose It?........ 189
12.4 Sample Design........ 190
12.5 Keywords and Notations........ 190
12.6 Advantages and Disadvantages of Sampling........ 191
12.7 Nonrandom Errors and Non-sampling Errors........ 192
12.8 Random Errors and Sampling Errors 192
12.9 Types of Samples 192
12.9.1 Probability Sample 192
12.9.2 Nonprobability Sample........ 192
12.10 Random Sampling........ 193
12.10.1 Systematic Sampling 194
12.10.2 Stratified Sampling (P, N) 195
12.10.3 Multistage Sampling 196
12.11 Nonrandom Sampling Methods........ 197
12.11.1 Convenience Sampling 197
12.11.2 Purposive Sampling 197
12.11.3 Quota Sampling........ 197
12.11.4 Cluster Sampling 198
12.11.5 Sequential Sampling 199
12.12 Sampling Distributions........ 199
12.13 Need for Sampling Distribution........ 202
12.14 Standard Error for Different Situations........ 204
12.14.1 When the Population Size Is Infinite 204
12.14.2 When the Population Is Finite 205
12.14.3 Sampling Distribution Based on Sample Means 205
12.15 Point and Internal Estimation........ 206
12.15.1 Point Estimate 206
12.15.2 Properties of Good Point Estimators 208
12.16 Interval Estimate........ 208
12.17 Confidence Interval Estimation for Large Samples........ 209
12.18 Confidence Intervals for Difference between Means........ 212
12.19 Estimating a Population Proportion........ 213
12.20 Estimating the Interval Based on the Difference between Two Proportions 214
12.21 Confidence Interval Estimation for Small Sample........ 215
12.22 Determining the Sample Size 219
Exercise 12........ 222

13. Hypothesis Testing, Parametric Tests, Distribution Tests, and Tests of Significance 225

13.1 Introduction 225

13.2 Null Hypothesis (H_0) 225

13.3 Alternative Hypothesis (H_1) 225

13.4 Type I and Type II Errors 227

13.5 Meaning of Parametric and Nonparametric Test 228

13.5.1 Parametric Test 228

13.5.2 Nonparametric Test 228

13.6 Selection of Appropriate Test Statistic 228

13.7 Methodology of Statistical Testing 228

13.8 Test for a Specified Mean: Large Sample 229

13.9 Test for Equality of Two Populations: Large Sample 238

13.10 Test for Population Proportion: Large Sample 243

13.11 Test for Equality of Two Proportions: Large Samples 247

13.12 Test for Equality of Two Standard Deviations: Large Samples 252

13.13 Student's t-Distribution 253

13.14 Properties of t-Distribution 254

13.15 Test for the Specified Mean: Small Sample 255

13.16 Test for Equality of Two Population Means: Small Samples 260

13.17 Paired t-Test for Difference of Mean 267

13.18 Chi-square Distribution 270

13.18.1 Properties of Chi-square Distribution 271

13.18.2 Chi-square Test 271

13.18.3 Test for Goodness of Fit 272

13.18.4 Tests for Independence of Attributes 279

13.18.5 Whenever the Expected Frequencies of the Cell Entries Are Less Than 5 287

13.18.6 Test for a Specified Population Variance 290

13.19 Snedecor's F-Distribution 294

13.19.1 Properties of F-Distribution 295

13.19.2 Test for Difference of Two Populations' Variances 295

13.20 Analysis of Variance (ANOVA) 301

13.20.1 One-Way Classification 303

13.20.2 Two-Way Classification 308

Exercise 13 315

Appendix A: Answers to Exercise Problems 325

Exercise 4 325

Exercise 5 325

Exercise 6 326

Exercise 7 326

Exercise 8 327

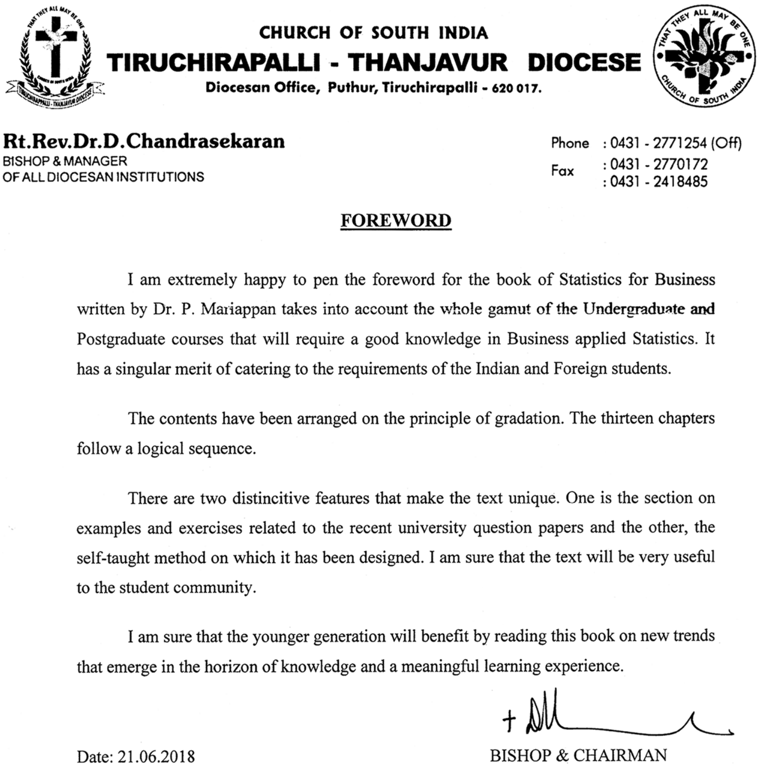
Exercise 9 328

Exercise 10 328
Exercise 11 328
Exercise 12 328
Exercise 13 330

Appendix B: ST Statistical Tables 335

Index 347

Foreword



CHURCH OF SOUTH INDIA

TIRUCHIRAPALLI - THANJAVUR DIOCESE

Diocesan Office, Puthur, Tiruchirapalli - 620 017.

Rt.Rev.Dr.D.Chandrasekaran
BISHOP & MANAGER
OF ALL DIOCESAN INSTITUTIONS

Phone : 0431 - 2771254 (Off)
Fax : 0431 - 2770172
: 0431 - 2418485

FOREWORD

I am extremely happy to pen the foreword for the book of Statistics for Business written by Dr. P. Mariappan takes into account the whole gamut of the Undergraduate and Postgraduate courses that will require a good knowledge in Business applied Statistics. It has a singular merit of catering to the requirements of the Indian and Foreign students.

The contents have been arranged on the principle of gradation. The thirteen chapters follow a logical sequence.

There are two distincitive features that make the text unique. One is the section on examples and exercises related to the recent university question papers and the other, the self-taught method on which it has been designed. I am sure that the text will be very useful to the student community.

I am sure that the younger generation will benefit by reading this book on new trends that emerge in the horizon of knowledge and a meaningful learning experience.

Date: 21.06.2018

BISHOP & CHAIRMAN

Preface

Statistics for Business has been contrived to serve as a textbook for students in business, computer science, bioengineering, environmental technology, and mathematics. In recent years, business statistics has been used widely for decision making in business endeavours. Like most tools, however, it is useless unless the user understands its application and purpose. To perform efficiently in the present complex world, the business executive ought to know enough about the basic principles of data analysis. To be certain that use is being made of all the information available to resolve a given problem, this text emphasises statistical applications, statistical model building, and determining the manual solution methods.

Target Audience

This textbook is meant to be used by beginners and advanced learners as a text in Statistics for Business or Applied Statistics. The following groups of students can benefit from using it:

- All undergraduate programmes having Applied Statistics at Major or Allied level.
- All postgraduate programmes having Applied Statistics as a course.
- The users of Applied Statistics who need a comprehensive reference.

Organisation

The text contains sufficient information for all the courses. This allows teachers ample flexibility in adopting the textbook to their individual class plans. The text includes an introduction to statistics and its business applications, data structures, data sources, and data collection; data representation; measures of central tendency; dispersion; skewness, moments, and kurtosis; correlation and regression analysis; probability; random variables and expectations; discrete probability distributions and continuous probability distribution; theory of sampling; and tests of hypothesis.

Highlights

- The problems discussed in the examples and in the exercises are related to the applied statistics paper of recent university examinations.
- This text is prepared based on ‘self-taught’ method.
- For most of the methods, the required algorithm is clearly explained using flow-charting methodology.

I do hope that this text will meet the needs of those for whom it has been actually designed.

Dr Perumal Mariappan

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Acknowledgements

Many people have played significant roles in the development and release of this text. My opinions on teaching Statistics for Business are greatly influenced by excellent teachers I had—great professors like Dr Karuppan Chetty, Prof. Genesis, Dr Subramaniam, and Dr A. Srinivasan just to name a few.

It is my pleasant duty to thank Rt. Rev. Dr D. Chandrasekaran, Bishop of Tiruchirappalli-Thanjavur Diocese, Trichy, for his wonderful Foreword to this text.

I record my earnest thanks to Prof. D. Swamiraj, Former Principal and Director, Bishop Heber College, Trichy, for his constant encouragement in writing textbooks.

My heartfelt thanks to my beloved Dr D. Paul Dhayabaran, Principal, Bishop Heber College, Trichy-17, for his constant encouragement and support, which gives me immense pleasure in bringing out many books like this.

I thank Professor P. Cruz, Retired Professor of English, Bishop Heber College, Trichy, for his tireless effort and patience in taking care of the language part of this text.

It is my duty to thank Miss. A. Jenifer Christinal and Mr. D. Surjith Jiji for their tireless effort in the proof reading of the full text.

I offer my special thanks to the helpful, competent CRC Press Publishing team.

I owe a great debt of thanks to the members of my family and friends who helped me in this achievement.

I am responsible for all the errors and would love to hear the suggestions and comments from the readers of this text.

I submit myself at the feat of the Almighty who always showers His blessings.

Dr Perumal Mariappan

Author

Dr Perumal Mariappan is a distinguished academician working as Head of the PG Department of Actuarial Science, Bishop Heber College, Trichy, since 1990. He has a PhD in Management Science and another in Applied Mathematics. Mariappan has more than twenty-five years of rich teaching experience in the field of Applied Mathematics, Statistics, and Business Administration in India and the United Arab Emirates.

Dr Mariappan received the best teacher award for the year 2004 given by The Association of Centre for Indian Intellectuals (CII), India. The prestigious 'Shiksha Rattan Puraskar' award given by the India International Friendship Society, New Delhi, was bestowed upon him in 2012. In 2015, he received several prestigious awards, including the Bharath Excellence Award given by the Friendship Forum, New Delhi; the 'Best Personalities of India' given by the Friendship Forum, New Delhi; the 'GRABS – Life Time Achievement Award' at Chennai; and the 'Abdul Kalam Life Time Achievement National Award' at GOA given by International Institute for Social and Economic Reforms, Bangalore. He received the Distinguished Teacher Award: Mathematics awarded by Management Teachers Consortium, Global Trust in Bangalore. Recently he received the 3E Scholar Best Teacher Award (2017) given by 3E Innovative Foundation, Delhi. Recently Mariappan received 'Adarsh Vidya Saraswathi Rashtriya Puraskar' Award and 'Best Teacher Award in the country' given by Global Management Council, Ahmadabad in 2017 and received the Best Faculty Award 2017–2018 given by Cognizant, Bangalore.[*]

Dr Mariappan's teaching and research interests include Business Mathematics, Business Statistics, Decision Sciences, Operations Management, Management Information Systems, Quantitative Methods in Operations Management, Numerical Methods, Optimization Techniques, Differential Equations, Partial Differential Equations, and Computer Programming.

Dr Mariappan has published hundred articles in leading academic international journals such as *International Journal of Management and System*, *OPSEARCH*, and *Indian Journal of Pure and Applied Mathematics*. He also has written numerous textbooks and has conducted three international conferences and three international workshops.

He has also presented research articles at many international conferences held in Atlanta, Costa Rica, Singapore, Philippines, Thailand, Dubai, Malaysia, and Sri Lanka. He served as Jury in an International Panel for the IFI International Panel and Conference held at the Groupe ESC Rouen, France. He has a life membership in All India Association for Educational Research (AIAER), Bhubaneswar, India; United Writers Association (FUWAI), Chennai, Tamil Nadu, India; the Association of Mathematics Teachers of India (AMTI);

[*] Recently received the RULA [Research Under Literal Access] Life Time Achievement Award 2019 on 26th FEB 2019 at Trichy, Powered by World Research Council and United Medical Council.

the Association of Mathematics Teachers of India, Chennai, Tamil Nadu; Operational Research Society of India, Madurai Chapter; and Probability and Statistical Society of India, Tirupathi. He is also a member of the editorial board of many international journals. Currently, he is guiding research scholars for their doctoral degree in the field of Management Science and Applied Mathematics. He reviews articles for many leading national and international journals.

1

Introduction to Statistics and Its Business Applications

1.1 Introduction

The word 'statistics' is derived from the Greek word *statistik*. Its meaning is political state, and the derivation suggests its origin. Administration of the state required the collection and analysis of data regarding population and property for purposes of war and finance. Now there is hardly any field of social activity or scientific research that does not find statistics useful.

The term 'statistics' has two meanings: (1) statistical principles and methods and (2) statistical data that have been developed to handle the data. When census reports are taken, it is a lot of data regarding the population. They are 'statistics' in the first sense of the word. On the other hand, the methods of collecting the data, the way samples are chosen for measurement, the methods of classifying and tabulating the data collected, the methods of analysing them and correlating them, the methods of interpreting them, and so much more form 'the statistical methods'.

The five stages mentioned are called the phases of a statistical investigation.

Croxton and Cowden define statistics as, 'the collection, presentation, analysis and interpretation of numerical data'.

Principles involved in interpreting the data are forming valid conclusions by analysing the data.

According to Bowley, '[s]tatistics may be called the science of counting'.

As per Boddington, '[s]tatistics is the science of estimates and probabilities'.

Spiegel states that statistics is concerned with scientific methods for collecting, organising, summarising, presenting, and analysing data and drawing valid conclusions and making reasonable decisions based on such analysis.

This implies that statistics can be said to include the study of the following:

Methods of collecting statistical data can be directly by researchers through mail cards or indirectly from existing published sources.

The various methods used to evaluate the reliability of the data.

Sampling Methods:

- Methods of classifying the data usefully and logically based on quantity, quality, time, or geographical regions.
- Methods of presenting the data in the form of easily understood tables, graphs, and other diagrams.

- Methods of calculating average, measures of variation, skewness, correlation, or association to understand the basic characteristic of the data.
- Principles involved in interpreting the data that are useful in forming valid conclusions by analysing the data.
- Principles involved in forecasting on the basis of the existing data.

Classification of Statistics:

The study of statistics can be classified into two broad areas: Descriptive Statistics and Inferential Statistics:

Descriptive statistics: Descriptive statistics can be defined as a set of methods involving the collection, presentation, characterisation, and summarisation of a set of data by means of numerical descriptions.

Inferential statistics: Inferential statistics can be defined as a set of methods that allow estimation or testing of a characteristic or attribute of a population, or the making of a judgment or decision concerning a population based only on sample results.

1.2 Is Statistics a Science?

Science is an organised body of knowledge, and statistics is the science of making decisions in the face of uncertainty. But strictly speaking, statistics is not a science like the physical sciences. To quote Croxton and Cowden again, '[s]tatistics should not be thought of as a subject correlative with physics, chemistry, economics or sociology. Statistics is not a science; it is a scientific method'. Statistical methods are an indispensable tool for the research worker in all sciences, physical, biological, or social. Wherever there are numerical data, the methods of statistics are useful.

1.3 Application of Statistics in Business

Statistical methods are used in the collection, analysis, and interpretation of quantitative data. Though these methods are used in every area of scientific investigation, they are especially useful to economists and businessmen. In fact, there is no field in business in which statistics do not come in handy as a tool for efficient and effective management. Statistical application to business can be used:

- for locating a new plant.
- to formulate a sales program.
- to expand the existing plant.
- to add new processes.
- to add new products.
- to have a new policy, etc.

For example, while selecting a location for a new plant, a businessman takes the following things into consideration:

- the most suitable site
- climate and adequate available labour
- availability of raw materials, transport facilities, and markets
- the tax laws, prevailing wage rates, cost of living, political climate, etc.

To make a decision, the necessary data must be collected, and with the help of statistics, a decision can be made. Similarly, in the field of sales planning, statistics is useful. By studying population and income statistics and consumption pattern, a businessman makes a forecast of its sales and plans them accordingly. It helps in making decisions regarding the stock of raw material inventory, stock of semi-finished goods inventory, and finished goods inventory, etc. In recent years, market research has been used to the fullest extent to study the consumer behaviour towards a product, their habits, the effectiveness of the various types of trade channels, advertisement media, and the functioning of rival companies, etc. Statistics is also extremely useful in the field of production planning. Modern business makes use of statistics to test the efficiency and performance of their workers to decide the promotional aspect. The statistical data also help a businessman compare his performance to that of his competitors. Another field in which modern industry uses statistics effectively is quality control. Banking and insurance companies, as well as other investors, look to production and finance and share price statistics of various companies to decide their investment programs. Insurance companies use actuarial statistics for determining their premium amounts. On the whole, statistics is an important tool to doing successful business. Statistics consists of a set of tools whose proper use helps in decision making. These tools are used in many fields other than business, for example, in biology, agriculture, psychology, education, and so on.

1.3.1 The Phases of the Statistical Decision-Making Process

Industry and government statisticians generally divide their tasks into different phases. They are

- Study design
- Data collection
- Data analysis
- Action on results

1.3.1.1 Study Design Phase

- *Question definition:* The manager defines the question in terms of the business need for information.
- *Alternative strategies:* The statistician develops and specifies alternative procedures for sampling, data collection, and analysis.
- *Strategy evaluation:* The manager and the statistician evaluate the advantages and disadvantages of the feasible alternatives.
- *Strategy selection:* The manager selects a strategy on the basis of cost and the importance of the information to the organisation.

1.3.1.2 Data Collection

- *Sample design:* The statistician plans the sampling procedure based on work done in second stage of Study Design and the selection made in fourth stage of Study Design.
- *Measurement:* Observations are chosen and recorded in a form that facilitates analysis.

1.3.1.3 Data Analysis

- *Statistical analysis:* Statistical methods are used for estimation summarising.
- *Reliability assessment:* Measures of possible error in results are calculated.
- *Report generation:* The statistician reports the results to the decision makers.

1.3.1.4 Action on Results

- An action is taken by the management based on the results of the study.

1.4 Responsibility of the Decision Maker

Using statistics to solve problems in business requires the involvement of a number of different people. The person who knows the functional aspect of the problem is as important as the statistician or the researcher. The phases and steps discussed state the important responsibilities of the manger and the statistician. Sharing of responsibilities is vital for the statistical decision-making process (Table 1.1).

TABLE 1.1

Responsibility of the Decision Maker

Manager's Responsibilities	Phases of Steps	Statistician Responsibilities
Study Design Phase		
Define the problem	1	
	2	Develop alternative strategies
Evaluate strategy	3	Evaluate strategy
Select strategy and approve the study	4	
Data Collection Phase		
	1	Design sampling procedure
	2	Measure and record data
Data Analysis Phase		
	1	Analyse the data
	2	Determine reliability
	3	Communicate results
Action Phase		
Act on results	1	

1.5 Functions and Limitations of Statistics

1.5.1 Functions of Statistics

Statistical methods are a helpful device to understand the nature of any phenomenon, if the methods are used carefully.

- For example, statistics can simplify complex data. The marks of 5000 students in a college by themselves make little sense. But when averages are calculated and ratios, like mean marks, passing percentage, etc., are evaluated, they give us a good idea of the students' standards.
- In the same fashion, a diagram graphically describing the trend of sales or profits of a company gives us the level of functioning of the company. It can expand a person's experience and test the validity of conclusions that we form from such experience.
- Statistical methods can compare data and measure the relationship between two factors. For instance, the mere list of prices on a day has no significance. But if the same is compared with the prices of the previous year by index numbers, it is possible to know the price trend.
- With the help of statistical methods, one can also find out the relationship between rainfall and crop yield; money in circulation and the price level; vaccination and immunity to disease; and so much more.
- With the help of statistical methods, the laws of other sciences can be tested. That is, to verify if the demand for a commodity falls when its price rises, referred to as 'The law of demand', we use statistical data covering a number of commodities.
- In the same way, whether cancer results from smoking, if tuberculosis can be prevented by taking special medicines, if eye defects are to the result of heredity, whether ammonium sulphate increases production of crops, and so on can all be verified by using statistical methods.
- Moreover, statistical methods help in the formulation of government policies and business policies and in the evaluation of the achievement and progress by the country or company.

1.5.2 Limitations of Statistics

Statistical methods have their own limitations.

- Statistical methods cannot be used for individual items.

  They deal only with mass data and shed light on the characteristics of the entire group. We can know the average per-capita income of a country by statistical calculations. But we cannot know the extent of the misery of a pauper. The mean mark of a class does not reveal the intelligence of its best student.
- A single statistic cannot determine the value of a group. It should be confirmed by other statistics and evidences.

Just because a particular school has a higher percentage of passing students, one cannot conclude that its boys are more intelligent. One of the reasons may be they have stopped the below-average students from taking a final exam. In the same fashion, if two companies, say A and B, had the same profit this year, but the company A has had a higher profit last year and the other had a lower profit? This situation doesn't imply that company B is progressing and that company A is declining; this year's profit alone does not show it. To make any kind of conclusions based on statistical data, we should study their whole background and all the related data.

- Statistical methods can measure only quantitative data.

  They cannot measure nonquantitative facts like culture, friendship, health, skill, pessimism, or honesty. To evaluate certain qualitative items, we use related quantitative features such as age to measure youth, marks for intelligence, or income for prosperity.

- Statistical methods must be handled only by experts.

  Statistical methods are a double-edged weapon and must be handled only by experts. If anyone makes a decision with a lack of expertise in statistics, it may lead to a wrong conclusion.

1.6 Distrust of Statistics

Just because vested interests have misused statistics for selfish purposes and have been exposed later, people tend to distrust statistics. The popular distrust in statistics is generally expressed by the following remarks:

Statistics can prove anything. Statistics is like clay of which one can make a God or a devil as desired. In statistics, we give importance only to the figures irrespective of who prepared them and how they were prepared. This particular aspect is exploited by interested parties; statistics is misused and wrong inferences presented to the people. Occasionally, the statistical tool can be misused because of ignorance. In a common situation, the data set given is not going to be verified in the sense as to whether it is reliable. A table generated with false information will produce a wrong picture. When false figures are expressed very precisely, people believe them blindly. Statistics is abused when faulty generalisations are made. This results from a lack of knowledge in the field of statistical methods and also because of individual bias. It is common that if someone has come across a number of such wrong inferences, he or she tends to distrust all statistics.

Thus, statistics is capable of being misused if handled unscientifically. It is a useful tool but also a delicate tool. Like drugs, if used poorly, it may cause harmful results. To use statistics as a proper tool, make sure that the figures are properly collected, are suitable for the problem under investigation, the complete background of the data is known, and the inferences are logical.

1.7 Nature of Statistical Law

1.7.1 Law of Statistical Regularity

The study regarding a part of a population (sample) is possible, and we can estimate statistically the characteristics of the whole of it. It is a result of the occurrence of the regularity in life and nature. The number of times the faces are going to occur in an unbiased die out of 1000 trials will be approximately equal. To study the change in the wage rate of workers in the population of any country, it is not necessary to study the entire workers of that country. It is enough to study 25% of the population. Based on the outcome, one can estimate exactly the changes in the earnings of all factory workers. The part of the population (sample) should be selected in such a fashion that all factory workers are included in the study.

It can be concluded that from a very large population, a moderately large number of items is selected at random, and the sample selected is likely to have the characteristics of the entire population from which the sample is selected. This is known as the Principle of Statistical Regularity. The concept of sampling exists based on this law. It also helps in making estimates for the future.

1.7.2 Law of Inertia of Large Numbers

The principle of large numbers is based on a similar reasoning as the principle of statistical regularity. Regarding coin tossing, if we toss the coin 3 times, we may get 3 heads or even 3 tails. If we do the experiment for larger number times, say, one million, nearly half will be heads and half tails. This indicates that the large numbers are more stable than smaller numbers. This clearly indicates that if the sample is bigger in size, the study results will be closer to the actual results of the population.

In statistics, inferences and forecasts are made because of the validity of the aforementioned laws. Occasionally, if the forecast is wrong, and it may be due to insufficient sample size.

Exercise 1

1. Define the term 'statistics'.
2. Explain the business applications of statistics.
3. Comment on the statement: 'Statistics can prove anything'.
4. State the limitations of statistics.
5. Why is statistics essential?
6. Comment on the statement: 'Statistics cannot be viewed as science'.
7. Explain the principle of statistical regularity and the principle of large numbers and their importance in sampling.

2

Data Structures, Data Sources, and Data Collection

2.1 Introduction

Data is a word of Latin etymology used to describe a collection of natural phenomena descriptors, including the results of experience, observation, or experiment, a set of premises or information within a computer system. This may consist of numbers, words, or images, particularly as measurements or observations of a set of variables. Experimental data are data generated within the context of a scientific investigation. Mathematically, data can be grouped in many ways.

2.2 Data Structures

A data set of some basic measurement or measurements of individual items is called elementary units, which may refer to people, households, firms, cities, TV sets, and so on. The same piece or pieces of information are recorded for each one. A piece of information recorded for every item (its cost, etc.) is called a 'variable'. The data set can be classified in three ways. They are as follows:

- By the number of variables (univariate, bivariate, or multivariate)
- By the kind of information (numbers or categories) represented by each variable
- By whether the data set is a time sequence or comprises cross-sectional data.

The complexity of the data set is decided based on the number of variables or pieces of information recorded for each item, and this will guide us to select the proper tool for analysis. That is, one must decide whether the number of variables present is univariate, bivariate, or multivariate data.

2.2.1 Univariate Data

Univariate (meaning 1 variable) data sets have a single piece of information recorded for each item. The basic properties of this single piece of information can be summarised using the statistical methods available.

Examples:

The statistical analysis of data collected regarding the income level through a marketing survey would reveal the distribution of incomes, specific income level, and variation in the income level and the number of people within any given range of income.

Statistical analysis of the quality control regarding production could be used to keep check on quality and to verify whether the production is carried over in a proper direction.

The statistical analysis regarding the bond ratings of the firms in an investment portfolio would indicate the risk of the portfolio.

2.2.2 Bivariate Data

Bivariate (meaning 2 variables) data sets have exactly 2 pieces of information recorded for every item. Application of statistical analysis would reveal the relationship between the 2 variables under study. Apart from this, the study would help us to predict the value of 1 variable when the value of the other variable is given.

Example:

Consider a table that comprises the cost of production per unit of different companies and the number of units produced of a specific commodity for the past 6 months. One of the bivariate statistical analysis tools, namely correlation analysis, can be applied to study the degree of relationship between the cost and the number of units produced. Also, using the regression analysis tool, we can estimate the cost of producing an item if we know the number of units to be produced and vice versa. Here in this analysis the cost of production is taken to be the first variable and the units produced are considered to be the second variable.

2.2.3 Multivariate Data

Multivariate data (meaning at minimum, 3 variables) sets have a minimum of 3 pieces of information recorded for every item. Statistical analysis can be applied to study the interrelationship among all the variables. Moreover, an estimation analysis can also be done by combining all the variables put together.

Examples:

1. Consider a table consisting of the information regarding gender, total years of experience, designation, performance level, and salary received record for each employee. Multivariate analysis could help us decide whether women are discriminated in terms of salary.
2. Consider a table that comprises the growth rate, strategy adopted, type of equipment used, investment level, and management style for each of several new firms. The statistical analysis would give a clear picture that out of all the information provided which combinations have been successful.

2.3 Data Sources

Business data are categorised into the following types: primary, secondary, internal, and external.

2.3.1 Primary Sources

A set of data collected by an individual or organisation directly from the field of inquiry for a specific purpose is called 'primary data'. These data are original in nature and collected by trained investigators. Most often, the data that are considered primary are published in some form by the collecting agent (e.g., government, civil bodies, trade associations, and so on). It is often the case that published primary data contain information on how the data were collected, along with suggested interpretations and uses of the data. The collection of primary data is not simple; it is tedious, time consuming, and costly.

2.3.2 Secondary Sources

Secondary data are essentially republished information. That is, if the same set of data is called 'primary' when it is in the hands of individuals or organisation, who collected directly from the field, the same will be called 'secondary' if it is in the hands of another person who is going to refer to the same for study. For example, the consumer price index is republished in the *Economic Times* and most major daily newspapers. These are usually characterised by the lack of information on how the data were collected and the dearth of suggested uses and interpretations. The important sources of the secondary data are publications of state and central governments, international bodies (UNESCO, UNI, etc.), foreign governments, trade associations, cooperative societies, labour and trade union reports, research papers published by the research scholars, and so on.

2.3.3 Internal Source

A set of data is considered to be internal data if it is obtained from well within an organisation and relates to the operations of the organisation. A set of data may be partially or fully available from an internal source, such as an organisation's computerised file containing sales figures, financial data, operating information, and so on.

2.3.4 External Source

A set of data is considered to be external data if it is collected from outside the organisation. This type of information may be available in the published financial periodicals, or it may be stored in an internal computer data bank accessible by an online computer terminal.

Advantages and disadvantages of primary data over the secondary data

- Primary data give the complete information about each data, whereas it is not always possible to get complete information from the secondary data.
- Secondary data may contain errors, but the primary data is error free.

- Exact definition and scope of the primary data are explicitly stated, but the same cannot be expected from the secondary data.
- The limitation of the primary data can be evaluated based on the method and mode of collecting it, but this facility is not possible in the case of secondary data.
- Primary data are collected by the researcher or the organisation directly from the field of study, but the secondary data are collected by somebody and provided for comparison purpose.
- Because the primary data are collected by the person concerned, it is more suitable for the study than the secondary data.
- Primary data is tailor made, but the secondary data is not.
- Primary data is more reliable than the secondary data.
- Collection of primary data needs more time than that of the secondary data.
- Collection of primary data is costlier than that of the secondary data.

2.4 Data Collection Inquiries

Statistical data are collected through statistical inquiries. These inquiries should be planned carefully, and the required data must be collected. After collecting, the data are classified and tabulated, analysed, interpreted, and presented in an easily understandable form. Careful planning is advocated for the success of data collection. Planning includes the clarity of its object and scope, the selection of the method of inquiry, and the degree of accuracy needed.

First, the object and scope of the inquiry should be predetermined carefully. This is because if there is no clear vision regarding the data needed, unimportant data may be collected and important data may be omitted. This may lead to waste of time and waste of money. For example, a cosmetic manufacturer, who wishes to know more details about the retail sales for his product. He may inquire about the number of ladies in the family of the consumer, adults and children; the number of items consumed; what brand of cosmetics they use; what type of quality they like, and so on. With this in view, the schedule for data collection is prepared.

Secondly, the methods of inquiry must be carefully selected. There are different methods of inquiry, and each is suited for a specific purpose. If the entire population has to be studied, a census inquiry is undertaken. If that is too costly or too time consuming, a sample study will be made. Particularly, if the sample is random and sufficiently large (20%–30% of the size of the population), the results will be quite good. After deciding the size of the sample, the next stage is in what fashion the data is going to be collected.

Primary data may be collected through the observation method or through a questionnaire method.

In observation method, the person who collects the data (investigator) asks no questions, but he observes carefully the phenomenon under consideration and records the essential data. Observation can be done by an individual or using mechanical device or electronic device. The major disadvantage of this method is the question of accuracy. That is, it is difficult to produce accurate data. The other issue is any physical difficulty on the part of the observer. It may cause inaccurate data. Owing to these difficulties, the questionnaire method is widely used for collecting the required data.

In the questionnaire method, the researcher designs a questionnaire that contains all the relevant questions needed for the study. The researcher gets the required answer from the respondents and accordingly records it. This method of collecting data can be conducted through personal interview, by mail, or by telephonic interview. In the personal interview method, the interviewer sits face to face along with the respondent and records responses. The only merit of the method is that it is more accurate and reliable. This is because the interviewer can clear up doubts and cross-check the respondents' answers. The disadvantages of this method are time consuming and costly. The cost and the time increase proportionately with the number of respondents.

In the mailing the questionnaire method, the questionnaire is mailed to the respondent's residential or official address. A cover letter requests that the respondents fill it out carefully and return it. This method is advisable if the respondents are spread over a wide geographical area and are literate. Accuracy and reliability are questionable because if any question is ambiguous or hard to understand, respondents may not answer the question correctly. Further, there is no guarantee that 100% of the sample will return the filled-in questionnaire. So, for an approximate 1:3 ratio of return, it is costly.

In the telephone interview method, the researcher asks the respondents relevant questions over the telephone. It is less expensive. The data collected through this method is somewhat accurate. The main problem with this method is that the respondent should have a telephone facility and ample time to talk on the phone. There should also be a restriction regarding the number of questions to be asked.

Among these methods, the questionnaire method is an efficient method, and the data can be collected very fast. It has a major restriction on certain sensitive aspects such as income, age, or personal life details that the respondent may not be willing to share with the researcher.

Third, the units of measurement must be carefully defined. It helps to obtain uniformity in data and enables comparisons and the drawing of valid inferences.

Fourth, it is highly essential to decide the degree of accuracy to which the data are to be collected.

Finally, the preparation of the questionnaire plays a vital role. It should contain all the necessary questions, but not a long list of items. The questions should be clear and easy to understand. The questions must be arranged in a sequential order, and capable of providing the necessary and accurate data.

To decide the quality of the questionnaire, a sample study of the questionnaire, called a 'pilot study', can be done. Based on the study report, the questionnaire can be modified, if necessary, before going to the field for collecting the data.

2.4.1 Survey Design

A survey design includes designing a questionnaire, pretesting a questionnaire, and editing the primary data.

2.4.1.1 Questionnaire Design

The success of data collection totally depends on how efficiently and imaginatively the questionnaire has been designed. Certainly, a defective questionnaire will never be able to collect the relevant data. The following points must be carefully considered while constructing the questionnaire.

Letter of introduction: A letter of introduction should be attached along with each questionnaire. It should specify the purpose of the study and should give assurance to the respondents about the maintenance of confidentiality. It must be designed in such a fashion that it should motivate the respondents to give good responses. It should provide a sense of satisfaction for the respondent.

Number of questions: There is a close relationship between the number of questions asked in the questionnaire and the satisfaction of the respondents. Hence, the number of questions should be limited to a few only. It helps the respondent give accurate answers. Too many questions may stress and strain the respondent. In turn, it will affect the accuracy of the data. Research states that the number of questions in the questionnaire should be between 20 and 50. If it is more than 20, try to put all the questions under proper subheadings to have clarity.

Structure of the questions: The questions should be simple, short, and easy to understand. It can be of 'yes' or 'no' type or multiple choice. The questions should be complete in all respects.

Nature of questions: The nature of the question should not be sensitive. Also, it should not ask personal or confidential information. If such confidential information is required, then a confidentiality agreement should be given to the respondents. The questions are to be designed in such a way that the answer does not require any kind of calculations.

Sequence of the questions: The questions should be arranged in a proper sequence in such a way that there is a continuity of responses, and it is not necessary for the respondent to refer to previous questions. It should be a mixture of introductory questions, crucial questions and light questions. Then only the respondent will get satisfaction.

Questions of cross verification type: The questionnaire should contain some questions that are going to help the reliability of the information provided by the respondent.

Uniqueness: Each question must be tested for its unique meaning, that is the questions must be designed to give the same meaning to each respondent. If it is left ambiguous, then the respondent may give different answers. Certainly it will mislead. So the clarity of the question must be tested carefully. If any change is needed, it should be carried out properly.

Markings for clarity: If any question needs clarity for answering, some sort of clarification can be given by means of an example towards the end. For this, the researcher can use footnotes.

2.4.2 Pilot Survey of the Questionnaire

After completing the design of the questionnaire, it must be pretested. This process is referred to as pilot survey. Obviously, this process precedes the actual survey work.

Pretesting the questionnaire permits the researcher to rectify the problems, inconsistencies, repetitions, and so on. The outcome of the pretesting needs to ensure that modifications to the existing questionnaire are made and must be carried out immediately before getting into the actual data collection from the respondents.

2.4.3 Editing Primary Data

After the process of data collection, the data must be edited before being analysed. The collected data must be verified to ensure completeness, consistency, and accuracy.

Completeness: Each questionnaire should be verified to see whether the respondents have answered all the questions. If any question(s) is found unanswered, try to contact the respondent to get the answer. If it is not possible to get the answer, drop that questionnaire for further analysis.

Consistency: Check each questionnaire carefully to see if there is any contradiction. If any contradiction is there, try to contact the respondent and get the answer clarified. If any modification is there, modify. If it is not possible to sort out the contradiction, drop that questionnaire.

Accuracy: The collected data must be verified for its accuracy. Even if thought it is not an easy job for the researcher, it has to be carried out carefully. If the inaccurate data are included, the outcome won't be accurate. For this, the researcher can use the random verification of the collected data.

2.4.4 Possible Errors in Secondary Data

Normally, there is more chance for the secondary data to contain errors. Hence, the user of the secondary data should be careful in employing the it. The errors can be categorised into transcribing errors, estimating errors, and errors due to bias.

Transcribing error: There is a chance of occurrence of an error while transcribing the secondary data.

Estimating error: A majority of the published secondary data may be predicted using the statistical estimation analysis. The conclusions should not be drawn by treating the secondary data as the reliable source.

Errors due to bias: Sometimes the secondary data set may contain assumed figures incorporated as a result of the natural bias of the estimator.

Points to be considered while using secondary data: Because of the negative factors, the users of secondary data should be careful about the following points and should decide how far this data set is useful for the study under consideration"

- the complete history about the data;
- the methods used for the collection of data;
- the time frame and the area covered;
- the source of reliability and the authenticity of the primary investigator; and
- the unitization of the measurements of the data collected.

The secondary data must be verified before using it. The user should not accept it based on its face value because there may be bias present, the size of the sample was small, there were computational errors, and so on. Hence, the user of the secondary data should take extra care while using the secondary data.

- The user should assure that the data collected is suitable for the problem under study. Suitability of the data can be decided based on comparing the nature and scope of the study.
- The most important factor is the reliability of the secondary data. This is necessary because the secondary data has been collected by somebody for a different study purpose. The researcher should confirm that the organisation that collected the data is unbiased. So, careful examination should be made before using the same.
- Before using the secondary data, it must be tested for its adequacy. That is, the data must be verified for its limitations based on the current study. If it is suited exactly for the study, it can be used; otherwise it should not be considered.

2.4.5 Census and Sampling Methods

Primary data become necessary whenever the secondary data is not available. The primary data can be obtained either by census method or sampling method.

Census method: When the researcher collects data from each individual of the population, it is called the 'census method' or 'complete enumeration method'.

Advantages

- Information regarding each and every member in the population can be obtained.
- The information collected is more accurate.

Disadvantages

- It requires a lot of time and huge amount of money.

Sampling method: Unlike census method, if the researcher collects data from some of the members of the population, it is called the 'sampling method'. This method is used extensively.

Example Questionnaire:

Non Resident Expatriates (NREs) in United Arab Emirates and their level of satisfaction

Name:
Gender: Male Female
Marital Status: Unmarried Married Single
Qualification: <10 +2 UG PG Engg. Medicine
Nationality:
E-mail ID:

1. Occupation:
2. Nature of Organisation: Government Private
3. Number of years in Dubai:
   <2 <4 <6 <8 <10 >=10
4. Monthly Income (AED):
   <1000 <3000 <5000 <7000 <9000 <11000 >=11000
5. Approximate total monthly expense (AED):
   <500 <1000 <1500 <2000 <2500 <3000 <3500 <4000 >=4000
6. Please specify the most expensive component of your monthly expense:
7. Are you staying with your family? Yes No
8. Number of dependents: 1 2 3 4 5 >5
9. If you are married, is your spouse is working? Yes No
10. Do you have savings? Yes No
11. Mode of savings:
    Deposit in a bank Investing in shares and bonds Real estate Others specify ………

Education

12. What do you feel about the various educational services available in United Arab Emirates?

Education	Excellent	Good	Satisfactory	Not Satisfactory
High School				
Secondary School				
Higher Education (UG)				
Higher Education (PG)				
Professional Degree				

13. What is your opinion regarding tuition fees? Please specify…………..
14. Regarding the higher education of your children, you are interested in putting them in?
    Home country Dubai USA UK

Housing

15. Your view regarding the rent structure in Dubai.
    Costly Moderately costly Cheap
16. Are you really interested in house sharing? Yes No
17. Do you have any problem in getting a rental house?
18. Are you facing any problem regarding water, electricity, and gas? Yes No
    If 'yes', state the reason……………………..
19. Have you come across any problem with the landlord?
20. What is your opinion regarding the key money concept?
    ……………………………..

Transportation

21. Your mode of transportation: Own car Bus Taxi Other mode
22. Are you satisfied with the public transport facility provided by the Dubai government?
    Very much satisfied Satisfied Moderately satisfied Not satisfied Neutral
23. Are you satisfied with the taxi facility provided by the Dubai government?
    Very much satisfied Satisfied Moderately satisfied Not satisfied
24. If you own a car, what do you feel about the fuel cost?
25. Briefly comment on the traffic in Dubai.
26. Opinion regarding Dubai police:
    Friendly Very strict Having concern for the foreigners No consideration
27. What do you feel about the municipality fines?
28. What do you feel about the police fines?
29. How many times do you visit your home country in a year?
30. Which airlines do you choose?
31. State the motivational factor that pressurized you to stay in UAE.
32. Do you feel that UAE is like your second home? Yes No
33. Any other suggestions, opinions, or comments

Food Stuff

34. State your opinion regarding the facilities provided by the government.

Particulars	**Excellent**	**Good**	**Fair**	**Satisfactory**	**Not Satisfactory**
Availability of supermarkets					
Price of the food stuff					
Quality of food stuff					
Shopping Malls					
Price of the products					
Electronic Items					
Quality of the electronic item					
Price level of the electronic items					
Availability of healthcare centres					
Cost level of the healthcare centres					
Quality services provided by the healthcare centres					
Police Services					
Individuals safety					
Municipality services					
Banking facilities					
Telecom services					
Insurance facilities					
Recreation facilities					

Visa Rules and Regulations

35. What do you feel about the rules and regulations of getting visa?

Type of Visa	Excellent	Good	Fair	Satisfactory	Not Satisfactory
Visit visa (family members)					
Visit visa (for friends)					
Resident visa					

Exercise 2

1. Explain the different kinds of data.
2. Discuss the different sources of collecting data.
3. State the advantages and disadvantages of primary data over the secondary data.
4. Explain the survey design.
5. Discuss the process of editing the primary data.
6. Comment on the statement: ‘Possible errors in secondary data’.

3

Data Presentation

3.1 Introduction

The successful use of the data collected depends on the way in which it is arranged, displayed, and summarised. This chapter covers the presentation and condensation of data.

3.2 Classification of Data

Classification is the process of arranging the data based on the similarities and dissimilarities. It is sorting.

3.2.1 Types of Classification

The data can be classified into four types.

1. Geographical Classification
   In this type of classification, the data is classified based on the area or region.

   Examples:
   - Production of rice state wise
   - Population of India state wise
2. Chronological Classification
   In this type of classification, the data is classified according to the time of its occurrence.

   Examples:
   - Production of a company can be represented based on week, month, or year
   - Sales data of a company for the past 5 years
   - Statistical data classified according to this type is called time series
3. Qualitative Classification
   Classification of data made to some nonmeasurable characteristics such as nationality, religion, employment, sex, and such ass is known as qualitative classification.

Examples:

Classification of population based on employment.

Classification of population based on mother tongue.

4. Quantitative Classification

   Classification of data according to some measurable characteristics is known as quantitative classification.

   Example:

   The employees of an organisation can be classified based on their monthly salary.

3.3 Data Presentation

The collected data can be presented in three forms.

3.3.1 Textual Form

The descriptive presentation of data is referred to as textual form.

In the 2004 Parliament election, only 60% of the Tamil Nadu people cast their votes.

This form of representation has the following demerits:

- It is too lengthy;
- Comparison of the data cannot be done at a glance; and
- The researcher may find it difficult to provide an appropriate conclusion.

3.3.2 Tabular Form

Tabular form is the arrangement of individual items into condensed form. This is the first step in statistical analysis. Tabulating statistical data consists essentially of grouping similar items into classes and summarising each group, usually by counting the number of items for each class. A complete table includes titles, headings, body, and sources, all of which clarify the full meaning of the data presented.

Title: The title describes the contents of the table. It should be placed at the top of the table. Normally, each table should contain table numbers, which will be useful for future reference.

Headings: Each column should have a caption called 'headings'. The headings must be selected based on the content of the column.

Body: Body refers to the numerical information with reference to the descriptions of the rows and columns given under different headings.

Source: The source must be placed at the end of the table. It contains the information from which the actual data is collected along with the year of reference.

Characteristics of a Good Table

- The title and the headings of the table must be simple and precise. Both should convey the meaning of the content of the table.
- Each table should bear a table number. The number can be like 1, 2, and so on or double-numbered, such as 1.1, 1.2. The first kind of numbers conveys the sequence of the table. The second kind of numbers conveys the sequence and which chapter or part of the material the table can be found.
- The use of the contents should be clearly mentioned (e.g., instead of a head, sales, it can be, sales in rupees).
- As far as possible, the length and breadth of the table must be evenly spaced.
- At times, it can be adjusted based on the lengthy columns.
- The columns to be compared must be placed in succession.
- The total values of columns must be placed at the bottom of the table.
- The source of the data should be mentioned at the end of the table.

3.4 Types of Variables and Data

Variables can be classified into two types. They are discrete and continuous variables.

Discrete Variable

A variable that can take only isolated or discrete value is called a 'discrete variable'.

Example:

X refers to the age of 5 members.

X (years)	14	17	19	20	21

Continuous Variable

A variable that assumes any real value (i.e., integer/fraction) within a specified limit is called a 'continuous variable'.

Example:

X refers to the range of marks secured by students.

X	0–10	10–20	20–30	30–40	40–50

Discrete Data

The values taken by the discrete variable is called 'discrete data'.

Continuous Data

The values taken by the continuous variable is called 'continuous data'.

3.5 Levels of Measurement

In statistics, measurement is the assignment of numbers to attributes of objects or observations. The level of measurement is a function of the rules used to assign numbers and is an important aspect in determining what type of statistical analysis can be approximately applied to the data.

3.5.1 Nominal Scale

The lowest or weakest level of measurement is the use of numbers to classify observations into mutually exclusive classes or groups. These observations are known as 'nominal data'.

Examples:

Sex of the employee

Male – 1 designates male	Female – 2 designates female

Parts produced by a machine

Effective – 1	Defective – 2

Weight

Overweight – 1	Normal weight – 2	Underweight – 3

It is also referred to categorical data. This is because the nominal variables identify categories of observations.

3.5.2 Ordinal Scale

When observations are ranked so that each category is distinct and stands in some definite relationship to each of the other categories, the data are called 'ordinal data'.

Example:

Product by people

*Good quality product	*Average quality product	*Poor quality product

3.5.3 Interval Scale

When the exact distance between any two numbers on the scale is known and when the data meet all the other requirements of ordinal data, they can be measured as interval data.

Example:

Two measures of temperature like Celsius and Fahrenheit.

$$F = (9/5) * C + 32$$

The 0 point for each scale is different temperatures, and the unit measure is different for each, but there exists a fixed relationship.

3.5.4 Ratio Scale

When measurements, having all the characteristics of the interval scale, also have a true 0 point, they have attained the highest level of measurement and are called ‘ratio data’.

Examples:

- The weight of an object may be measured in grams, ounces, or other measures.
- The origin of each item is the same and is 0 weight.
- The distance between 2 places may be measured in miles and kilometres.

3.6 Frequency

Number of times a value repeats itself is called the ‘frequency’. Usually, it is denoted by the letter ‘*f*’.

Example:

X	4	5
f	5	1

It implies that the value 4 occurs 5 times and the value 5 occurs only 1 time.

3.6.1 Frequency Distributions

Problems occur when data to be entered represents the number of items in each class. This type of classification is called a ‘frequency distribution’. When the variable counted is not a nominal variable, there may be problems with the definition of classes. The main considerations in constructing a frequency distribution are:

- Determining the number of classes.
- Deciding the size of the classes.

Number of Classes

A frequency distribution must be made with suitable number of classes. If the classes are few, the original data will be compressed. Each class will be crowded, and the information may be lost. If there are too many classes, many of them will contain only a few frequencies. The distribution will look irregular. Based on research, distribution is optimized if the total class intervals are between 6 and 15.

Example:

Suppose the marks secured by 50 students in a class are given and they range from 0 to 100, then the number of classes can be decided as 10.

Size of Classes

As far as possible, all classes should be of the same size. To decide the size, find the range (max – min) and divide it by the number of intervals.

NOTE: When the items being classified contain a few extremely large or small items, it is usually impossible to set up equal class intervals.

Class size = ([maximum value – minimum value]/number of class intervals)

Example:

Consider two class intervals 10–20 and 20–30. In the class 10–20, 10 is the lower limit and 20 is the upper limit.

Class width = upper limit – lower limit = 20 – 10 = 10

Midpoint of the class:

Midpoint of the class interval = (lower limit + upper limit)/2

In this example,

Class Interval	Midpoint
10–20	(20 + 10)/2 = 15
20–30	(30 + 20)/2 = 25

3.7 Types of Class Interval

The types of class interval can be classified into three types. They are:

- Exclusive method,
- Inclusive method, and
- Open-end method.

Exclusive Method

In this type, the class intervals are arranged in such a way that the upper limit of a class is the lower limit of the subsequent class. Here the class intervals are continuous.

Example:

Monthly Income (Rs.)	Number of Employees
500–1500	30
1500–2500	10
2500–3500	20
3500–4500	30
4500–5500	10

In this exclusive method, the extreme values will be always included in the successive interval. Here the class intervals are continuous.

Inclusive Method

In this method, both the lower and the upper limits are included in the same interval.

Example:

Monthly Income ($)	Number of Employees
500–1499	5
1500–2499	4
2500–3499	3
3500–4499	4
4500–5499	2

The given discontinuous interval can be converted into a continuous interval. This is possible if the difference between the lower limit of the succeeding interval and the upper limit of previous interval are same.

Find, d = ½ * (lower limit of the succeeding interval – upper limit of the previous interval)

Subtract the value of d from the lower limits and add d to the upper limits.

$$d = ½\ (1500 - 1499) = 0.5$$

Modified table based on the income of the employees.

Monthly Income ($)	Number of Employees
499.5–1499.5	5
1499.5–2499.5	4
2499.5–3499.5	3
3499.5–4499.5	4
4499.5–5499.5	2

Open-End Method

In this method, the lower limit of the initial class and the upper limit of the end class will not be given. This is possible, when there is a large gap between the minimum and the maximum values.

Example:

Income Level (Rs.)	Number of Employees
<1000	10
1000–2000	10
2000–3000	20
3000–4000	10
>4000	10

NOTE: Apart from these two, if the other classes are having uniform length, convert the first and the last classes at the same width.

<1000 _____0–1000 and
>4000 _____ 4000–5000.

3.8 Tally Mark

- Tally mark is used to count the number of times a particular value of the variable is repeated.
- The tally mark is denoted by the symbol ‘\’.
- To count the total tally marks against a variable is done easily. Tally marks are prepared in the form of blocks; each block contains five tally marks.
- To ascertain the block, put every fifth tally mark in a cross position.
- The total frequency of the variable is the count of tally marks against the variable.

3.9 Construction of a Discrete Frequency Distribution

Example:

To explain the process, consider the sample study in which 25 employees were surveyed to find the number of members in their family; the data obtained are

4	2	6	5	3	3	4	3	2	2	2	1	2
4	4	1	3	1	3	5	6	2	6	2	3	

The data given can be condensed using discrete frequency distribution.

Number of Members	Tally Marks	Frequency
1	\ \	2
2	𝍸 \ \ \	8
3	𝍸 \	6
4	\ \ \ \	4
5	\ \	2
6	\ \ \	3
	Total	25

3.10 Construction of a Continuous Frequency Distribution

Example:

Construct the continuous frequency table

To explain the process, consider the sample study in which 24 employees of a company were surveyed to find their income. The data obtained are:

1800	1250	1760	3500	6000	2500
2700	3600	3850	6600	3000	1500
4500	4400	3700	1900	1850	3750
6500	6800	5300	2700	4370	3300

Step 1:
Find the minimum and the maximum values.

Minimum value = 1250; Maximum value = 6800

Range = max – min == 6800 – 1250 = 5550

Number of class intervals = 5550/1000 = 5.55 = 6 approximately

Income ($)	**Tally Marks**	**Frequency**
1000–2000	𝍸 \	6
2000–3000	\\\	3
3000–4000	𝍸 \\	7
4000–5000	\\\	3
5000–6000	\	1
6000–7000	\\\\	4
	Total	24

Note: Tally mark for 6000 should be made in 5000–6000.

3.11 Cumulative and Relative Frequencies

Apart from the regular representation of frequency against each class, it is possible to calculate their cumulative frequency or relative frequency or both.

Cumulative frequency: The cumulative frequency of a class interval is the summation of all frequencies up to that class interval for which the cumulative frequency is needed.

Relative frequency: The relative frequency of the classes can be obtained by dividing the actual frequency of the class by the total frequency. It shows the percentage for each class. It helps to understand the concept of probability and to compare 2 or more sets of data.

Example:

Form the frequency distribution for the following data given weights in pounds of 30 college students: Take class intervals of 10 units each.

132	104	166	143	129	119	108	151	111	134
114	155	138	131	124	157	130	132	132	145
122	121	148	144	139	147	135	126	136	142

Also, construct the cumulative frequency and relative frequency distribution.

Given the length of the class interval = 10

Minimum weight = 104; Maximum weight = 168

Range = 166 − 104 = 62

Number of intervals = range/length = 62/10 = 6.2 = 7 approximately

Weight (lbs.)	**Tally Marks**	**Frequency**	**Cumulative Frequency**	**Relative Frequency**
100–110	\\	2	2	2/30 = 0.067
110–120	\\\	3	2 + 3 = 5	3/30 = 0.100
120–130	𝍸	5	5 + 5 = 10	5/30 = 0.167
130–140	𝍸 𝍸	10	10 + 10 = 20	10/30 = 0.333
140–150	𝍸 \	6	20 + 6 = 26	6/30 = 0.200
150–160	\\\	3	26 + 3 = 29	3/30 = 0.100
160–170	\	1	29 + 1 = 30	1/30 = 0.033
	Total	30		1.000

The cumulative frequency column helps find how many elements are up to that class interval. In the preceding example, there are 5 elements up to 120 pounds. The relative frequency column helps to find the percentage of elements present in that interval. In the example, 20% items fall in the interval (150–160) pounds.

3.12 Diagrammatic Representation of Data

Usually the statistical data can be presented in the form of statements and tables; additionally, it can also be represented in the form of diagrams. Diagrams are ideal visual methods of presenting data. It helps people understand the data easily. Engineers, managers, technical men, and businessmen have been using them for a long time. What is hidden in a mass of data is brought out clearly and within a second, we get a cross-sectional diagram of the whole situation. A diagram of daily or weekly sales tells the manager quickly the trend of the business. Government sectors also use the diagram to show the nation’s economic development. It can be made attractive with the help of the advancements in computers and can make a colourful diagram. One of the main functions of the statistical method is to present complex data in a simple and comparable form. As the diagram does this job well, it is popular. As the saying goes, ‘one diagram is worth a thousand words’.

3.12.1 Advantages and Disadvantages of Diagrammatic Representation

Advantages

- It is used for interpolation and extrapolation of missing data.
- Using this, mode and median can be evaluated.
- Estimation analysis can be made.

Disadvantages

- Accuracy is not as high as compared with a table.
- There is the possibility of providing a wrong picture of the situation.
- It can give only limited information.
- Sometime diagrammatic representation reads approximation.
- It takes more time to prepare.

Despite these disadvantages, diagrams are popular, and properly prepared, they are remarkably useful.

3.12.2 Types of Diagrams

The following diagrams are used for graphical representation of data collected:

- Bar diagram
- Pie diagram
- Histogram
- Frequency polygon
- Frequency curve
- Line diagram and
- Ogive, curves, etc.

Each diagram should have a title. The title should answer the questions what, when, and where. It should be neat and not crowded. The axes must be spelt properly. To differentiate each group of items, different colours can be used. If different colours are used, then the reference regarding which colour refers to what must be shown.

3.12.2.1 Bar Diagram

The bar diagram is simple to draw and easy to read. It is widely used. It is useful for comparing simple magnitudes. It can be classified into simple bar or vertical bar, horizontal bar, multiple bar or compound bar, and component bar; and bilateral bars show profits and losses.

Consider the following points before preparing it

- Proper scale must be used.
- The bars should be of the same width.

- Uniform space must be given between the bars.
- Descriptions of the bars and components are usually given in the diagram itself.
- The title and the diagram number should be mentioned.

Example:

Number of cars sold by the Ford Company in the following months of 2017:

Month of the Year	No. of Cars Sold
January	1000
February	1100
March	1100
April	1200

Construct a bar diagram to represent the same.

Solution:

Consider the months along *x*-axis and the number of cars along *y*-axis. Erect four bars for each month based on the number of cars sold.

The diagram shown can be called a 'simple bar diagram' or 'vertical bar diagram'. If we change the setup so that the *x*-axis is number of cars sold and the *y*-axis shows different months, then the resulting diagram is referred to as a 'horizontal bar diagram' (Figure 3.1).

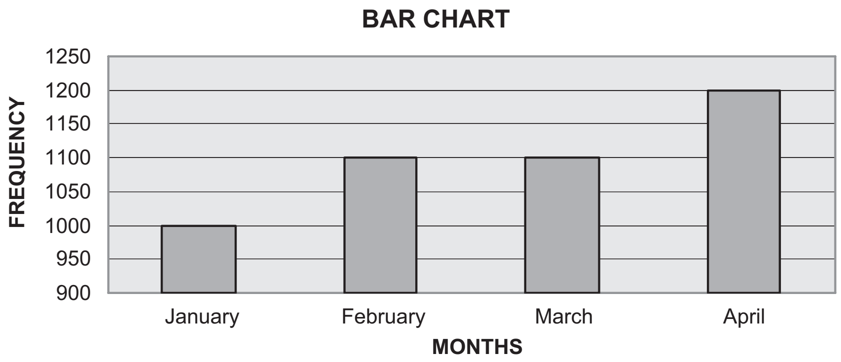


FIGURE 3.1
Bar chart.

Example:

Represent the following by suitable diagram:

Year	Export (in 1000 tonnes)	Import (in 1000 tonnes)
2000	4000	1500
2001	4500	2000
2002	5000	1000
2003	3000	2000

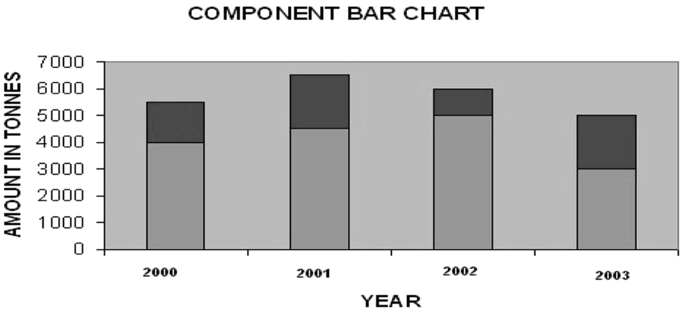


FIGURE 3.2
Component bar chart.

For this situation, draw a component bar or multiple bar diagram (Figure 3.2). Represent export and import side by side.

X-axis refers to the year, and *y*-axis refers to the units in 1000 tonnes.

Year	Export (in 1000 tonnes)	Import (in 1000 tonnes)	Total Value of Export and Import
2000	4000	1500	5500
2001	4500	2000	6500
2002	5000	1000	6000
2003	3000	2000	5000

Example:

Draw a deviational bar diagram for the following data of XYZ Company Ltd.

Year	Exports	Imports
1999	1500	1400
2000	2000	2200
2001	1700	1400
2002	2000	2100
2003	2500	2000

Construct the following table based on the difference (Export – Import)

Year	Exports	Imports	Balance (Positive)	Balance (Negative)
1999	1500	1400	100	—
2000	2000	2200	—	200
2001	1700	1400	300	—
2002	2000	2100	—	100
2003	2500	2000	500	—

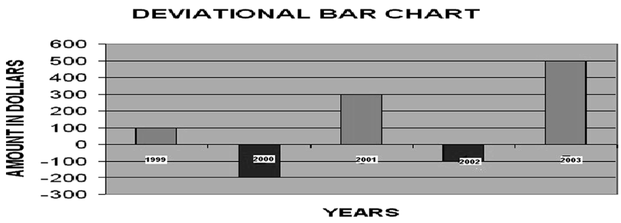


FIGURE 3.3
Deviational bar chart.

In the preceding table, the excess amount in exports is treated as positive, and the deficit amount in exports is treated as negative (Figure 3.3).

Also, this chart given is referred to as a 'deviational bar diagram'.

Example:

	Total Population (%)	
Particulars	**Villages**	**Towns**
Infants (I)	5	10
Young children (YC)	5	10
Boys and girls (BG)	15	20
Young men and women (YMW)	20	20
Middle-aged men and women (MMW)	30	20
Elderly person (EP)	20	25

Draw the multiple bar chart for the table of data (Figure 3.4).

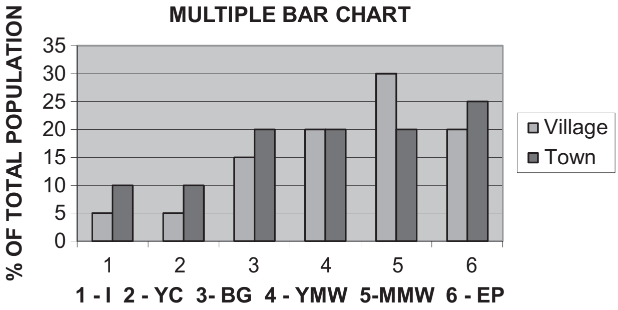


FIGURE 3.4
Multiple bar chart.

3.12.2.2 Pie Diagram

A pie diagram is a circular diagram. The component parts are shown as different sectors. The total figure is represented in a circle, and the total angle is taken as 360°, and for each component parts, the proportional angle is calculated. The desired degrees are marked off on the circumference and sectors are drawn to denote the parts. They are coloured differently and a description given therein or in a separate legend. A title is given as well. It is used to show the percentage change in the components of a total.

Working procedure

$$\text{Total percentage} = 100\%$$

$$\text{Total angle} = 360°$$

$$1\% = 360/100 = 3.6°$$

where 3.6° will represent 1% of the whole. For example, if 1 component is 10%, it implies that it is equivalent to 3.6° * 10 = 36°.

Example:

The following is an extract of the expenditure of the state government of Tamil Nadu on different heads: (1 unit is equivalent to 1 lakhs of dollars)

Different Heads	Expenditure (lakhs of $)
Direct demands on revenue (DDR)	3,00,000
Administration (ADMIN)	20,00,000
Other items (OI)	22,00,000
Overall expenses (Total)	45,00,000

Draw a pie chart (Figure 3.5).

Step 1: Express each component in percentage of the total expenditure.

$$\text{DDR} = (300000/4500000) * 100 = 6.7\%$$

$$\text{ADMIN} = (2000000/4500000) * 100 = 44.4\%$$

$$\text{OI} = (2200000/4500000) * 100 = 48.9\%$$

Step 2: Evaluate the equivalent component's degree.

$$\text{DDR} = 6.7 * 3.6 = 24.12;\ \text{ADMIN} = 44.4 * 3.6 = 159.84$$

$$\text{OI} = 48.9 * 3.6 = 176.04$$

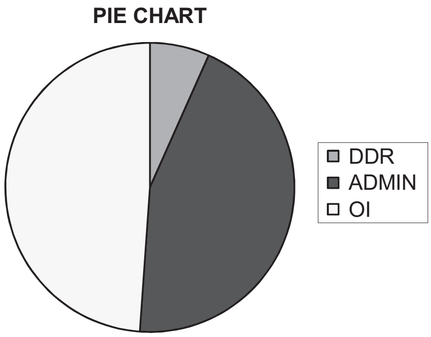


FIGURE 3.5
Pie chart.

Components	Amount (in lakhs of $)	Component (%)	Degrees of Component	Cumulative Degrees of Component
DDR	300000	6.7	24.12	24.12
ADMIN	2000000	44.4	159.84	183.86
OI	2200000	48.9	176.04	360
Total	4500000	100	360	—

Here the circle is separated into three segments.

3.12.2.3 Histogram, Frequency Polygon, and Frequency Curve

A histogram is the method of reporting a frequency distribution in the form of a graph. It consists of bars of the same width, each referring to class, and their heights referring to the class frequencies. Mark the midpoint of each bar on the top and move the midpoints of the preceding and succeeding classes of the initial class and the last class, respectively. Link all the midpoints using a straight line, and then the resultant graph is said to be a 'frequency polygon'. Link all the midpoints using smooth curve (free-bend), and then the resulting graph is said to be a 'frequency curve'. To draw the histogram, normally we take the class limits of the variable along the *x*-axis, and the frequencies of the class interval on the *y*-axis.

NOTE 1: If the class intervals are uniform in length and are not continuous, then first it must be converted into a continuous type of interval.

NOTE 2: If the class intervals do not having equal width, then the frequencies must be adjusted based on the width of the class interval.

Example:

Monthly sales (in lakhs of $)	10–20	20–30	30–40	40–50
Number of companies	3	4	2	1

Obviously, the lengths are uniform, and the intervals are continuous.

Histogram (Figure 3.6)

Step 1: Take the monthly sales along the *x*-axis and the number of companies along the *y*-axis.

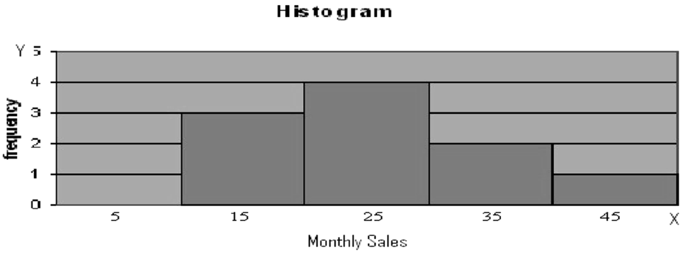


FIGURE 3.6
Histogram.

Step 2: On each class interval, erect a rectangle (of uniform length) with the height equal to the frequency of that class. If we proceed like this, then we get a series of rectangles.

3.12.2.4 Frequency Polygon

Select the midpoints of the intervals, including the preceding and succeeding class intervals. Link all those midpoints using a straight line. The resulting graph is the required frequency polygon (Figure 3.7).

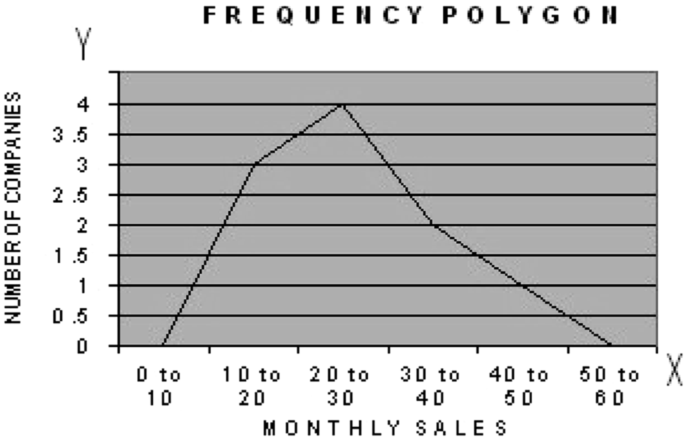


FIGURE 3.7
Frequency polygon.

3.12.2.5 Frequency Curve

Select the midpoints of the intervals, including the preceding and succeeding class intervals. Link all those midpoints using free hand. The resulting graph is the required frequency polygon (Figure 3.8).

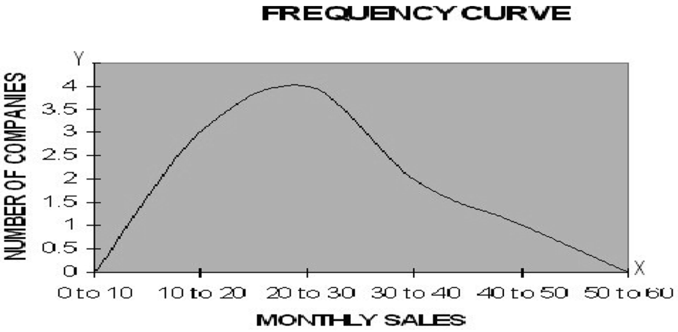


FIGURE 3.8
Frequency curve.

3.12.2.6 Ogive Curve

Ogive is a cumulative frequency curve. It can be evaluated in two ways as 'less than' or 'more than'. Two Ogive curves can be drawn from a given set of data. Both will intersect at a point. Always consider the cumulative frequency on the y-axis. To get less than Ogive curve plot (midpoint of the class interval, less than cumulative frequency), link all the points using a straight line. To get more than Ogive curve, plot (midpoint of the class interval, more than cumulative frequency), link all the points using a straight line (Figure 3.9).

Example:

Construct Ogive curves for the following data:

Weight (kg)	30–40	40–50	50–60	60–70	70–80
No. of students	400	500	700	300	100

Step 1: Construct the 'less than' and 'more than' cumulative frequency.

Lower Limit of the Class	Cumulative Frequency (less than <)	Cumulative Frequency (more than >)
30	0	2000
40	400	1600
50	900	1100
60	1600	400
70	1900	100
80	2000	0

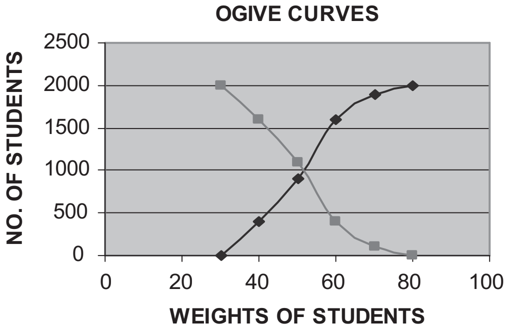


FIGURE 3.9
Ogive curve.

Take the lower limits of the class intervals on the x-axis and frequency on the y-axis. Draw the less-than Ogive curve by considering the points (lower limit, less-than frequency) and the more-than Ogive curve by considering the points (lower limit, more-than frequency).

3.12.2.7 Line Diagram

Among the given 2 items base entry and item A entry, consider the base entry on the x-axis and item A entry on the y-axis. Plot the coordinate points (base, item A) and link all the points using a straight line (Figure 3.10).

NOTE: If more than 1 item is given, draw the different lines using different colours (base, item1), (base, item2). Usually, the base entry may be year, month, name of companies, etc.

Example:

Present the following data graphically.

Year	Area (in lakhs of acres)	Production (in lakhs of tons)
2003	500	250
2004	550	275
2005	600	275
2006	650	300
2007	700	350

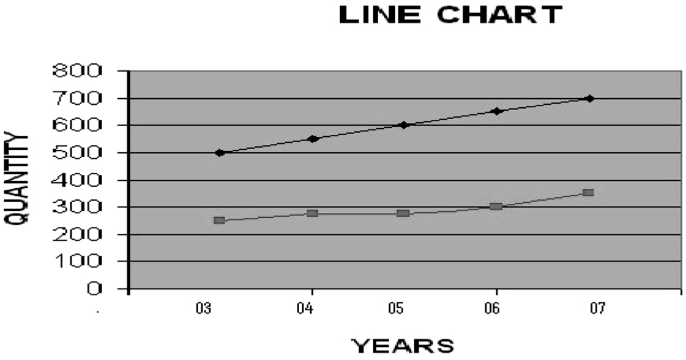


FIGURE 3.10
Line chart.

Exercise 3

1. Draw a histogram, frequency polygon, and frequency curve for the following frequency distribution:

Weekly wages (Rs.)	500–600	600–700	700–800	800–900	900–1000	1000–1100	1100–1200
No. of workers	15	25	10	10	15	20	5

2. Draw the bar chart to represent the following data related to a car manufacturing Industry.

Month (2007)	Jan	Feb	Mar	Apr	Jun	Jul
No. of cars produced	2000	2030	2100	2100	2200	2300

3. Draw the histogram, frequency polygon, and frequency curve for the following data:

   BMT, Inc., manufactures performance equipment for cars used in various types of racing. It has gathered the following information on the number of models of engines in different-size categories used in the racing market it serves:

Class (Engine size [in^3])	**Frequency (Number of Models)**
101–150	1
151–200	7
201–250	7
251–300	8
301–350	17
351–400	16
401–450	15
451–500	7

4. Draw a pie chart to represent the following data relating to the production cost of a manufacturer.

Cost of raw material	Rs. 2 lakhs
Cost of human resources	Rs. 3 lakhs
Other Over heads	Rs. 2 lakhs

5. Draw the histogram, frequency polygon, and frequency curve for the following data set referring a frequency distribution for the usage times of 50 automated teller machine (ATM) customers:

Time (s)	**Frequency**
20–25	1
25–30	7
30–35	10
35–40	9
40–45	9
45–50	6
50–55	5
55–60	3

6. Eighty packages have been randomly selected from a frozen food warehouse, and the age (n weeks) of each package is identified. Given the frequency distribution shown, draw the proper diagrammatic representation for the ages of the packages in the warehouse inventory.

Age (weeks)	**Number of Packages**
0–under 10	25
10–under 20	17
20–under 30	15
30–under 40	9
40–under 50	10
50–under 60	4

7. Draw the less-than and more-than Ogive curves for the following data.

Life (Years)	**No. of Refrigerators A**
0–2	5
2–4	16
4–6	13
6–8	7
8–10	5
10–12	4

8. Draw the multiple bar chart for the distribution of wages in two factories *X* and *Y*.

Wages ($)	**No. of Workers, *X***	**No. of Workers, *Y***
50–100	2	6
100–150	9	11
150–200	29	18
200–250	54	32
250–300	11	27
300–350	5	11

9. Frequency distribution showing the number of motorists in each speed category on a stretch of interstate highway. The distribution of data is as follows:

Speed (mph)	**Number of Motorists**
45–under 50	1
50–under 55	9
55–under 60	14
60–under 65	23
65–under 70	16
70–under 75	16
75–under 80	12
80–under 85	8
85–under 90	1

Draw the necessary diagram for the data.

4

Measures of Central Tendency (MCT)

4.1 Introduction

Statistical methods are needed for summarizing and describing the collected numerical data. The main objective of this chapter is to introduce one representative value that can be used to identify and summarize an entire set of data. This representative value is going to be helpful in making decisions based on the data collected. Measures of central tendency (MCT) are used to set the central value around which the data are spread over.

4.2 MCT

The average of a distribution is its representative size. As most of the items of the series cluster around the average, it is called a 'measure of central tendency'. The average is computed to reduce the complexity of the data. The entire distribution is reduced to one number, which can be considered typical of an important characteristic of the population, and the same can be used in making comparisons and in examining relations with other distributions.

For example, it is not possible to remember the individual's income in crores of earning people in India. By considering all the data related to income, if the average income is evaluated, we get a single value, which is going to represent the entire population. The commonly used averages are as follows:

- Arithmetic mean
- Median
- Mode
- Geometric mean
- Harmonic mean

4.2.1 Properties of Best Average

It should be

- Rigidly defined.
- Based on all observations of the series.

- Easy to calculate and simple to understand.
- Capable of further algebraic treatment.
- Free from the extreme values (i.e., it should not be affected by extreme values).

The arithmetic mean is ideal in these respects, even though the other averages are also useful in certain specific cases. The median is quite useful for studying data not capable of direct quantitative measurement like skin colour, etc. The measure mode is good when the extreme values are not well defined.

4.3 Arithmetic Mean

The evaluation of mean depends on the nature of data. The data for evaluation can be categorized into

- Discrete data (DD)
- Discrete data with frequency (DDF)
- Continuous data with frequency (CDF)

4.3.1 Discrete Data

Consider the given set of n number of discrete values X_i $[i = 1, 2, \dots , n]$
$X_1, X_2, X_3, \dots , X_n$. Then the arithmetic mean is defined as

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

4.3.2 Discrete Data with Frequency

Consider the given set of n discrete value corresponding with n different frequencies X_i and f_i $[i = 1, 2, \dots, n]$

X	X_1	X_2	X_3	…	…	X_n
f	f_1	f_2	f_3	…	…	f_n

Then the arithmetic mean is defined as

$$\overline{X} = \frac{\sum_{i=1}^{n} \{f_i \times X_i\}}{\sum_{i=1}^{n} f_i}$$

4.3.3 Continuous Data with Frequency

Consider the given set of data. Verify whether the given class interval is continuous. If not continuous, change it into a continuous one with proper methods. After that, verify whether the length of the class intervals is uniform. If not uniform, make proper adjustment of the variable length.

Class Interval	$L_1 - U_1$	$L_2 - U_2$	$L_3 - U_3$	…	$L_n - U_n$
f	f_1	f_2	f_3	…	f_n

where L_i is the lower limit of the *i*th class interval, and U_i is the upper limit of the *i*th class interval.

For each interval choose the mid-value and call it X_i (mid value of the *i*th class interval).

$$X_i = [L_i + U_i]/2$$

Mid-class Interval, X	X_1	X_2	X_3	…	…	X_n
f	f_1	f_2	f_3	…	…	f_n

Then, the arithmetic mean can be computed using the following relationship:

$$\overline{X} = \frac{\sum_{i=1}^{n} \{f_i \times X_i\}}{\sum_{i=1}^{n} f_i}$$

Alternative method

Mid-class Interval X	X_1	X_2	X_3	…	…	X_n
f	f_1	f_2	f_3	…	…	f_n
d	d_1	d_2	d_3	…	…	d_n

Find, $d_i = [X_i - A]/h$; for all $i = 1, 2, \dots, n$
A: any value of *X*, preferably the middle value.
h: the length of the class interval

$$\overline{X} = A + h \times \frac{\sum_{i=1}^{n} \{f_i \times d_i\}}{\sum_{i=1}^{n} f_i}$$

Relative advantages of arithmetic mean are listed as follows:

- easy to understand.
- easy to calculate.

- value is definite.
- familiar to all.
- based on all observations.
- least affected by fluctuations of sampling.
- stable, in the sampling sense.
- capable for further algebraic treatment.
- used to evaluate the mean deviation.
- used to evaluate the standard deviation.

4.4 Mathematical Properties of Arithmetic Mean

1. The sum of all deviations of the observations from the mean is 0. The same can be denoted mathematically, $\sum_{i=1}^{n}[X_i - \overline{X}] = 0$.

   Consider, $\sum_{i=1}^{n}[X_i - \overline{X}] = \sum_{i=1}^{n}[X_i] - \sum_{i=1}^{n}\overline{X} = n\,\overline{X} - n\,\overline{X} = 0$

   NOTE: All the evaluated statistics like mean, median, and mode are always constant.

2. The sum of the squared deviations of the observations from the mean is minimized. The same can be mathematically denoted as $\sum_{i=1}^{n}[X_i - \overline{X}\ ]^2$ is minimum.
3. Composite mean (combined mean) can be evaluated for any number of groups. If we know the means of different groups, then the composite mean can be evaluated. Let $(\overline{X}_1, n_1)$, $(\overline{X}_2, n_2)$, … , $(\overline{X}_n, n_n)$ be the mean and size of different groups, then the composite mean X can be evaluated using the relation

$$\overline{X} = \{(\overline{X}_1 \times n_1) + (\overline{X}_2 \times n_2) + \ldots + (\overline{X}_n \times n_n)\}/[n_1 + n_2 + \ldots + n_n]$$

Disadvantages of arithmetic mean related to other averages

1. If the number of items in a series is small, the more extreme items affect the arithmetic mean.

   Consider the marks secured by 6 students in mathematics

   100, 100, 100, 0, 0, and 0. The average mark is 50, which is not representative of the class.
2. In an organization, the production rate of 4 of its members per hour is A – 30, B – 30, C – 30, and D – 46. The average comes out to be 34; hence, the average 34 cannot be fixed as the production standard because many of the employees cannot achieve this.
3. The average cannot be evaluated even if 1 item in the given series is not known.

4. It cannot be located by mere inspection. It requires computation. In certain situation, the mean value found is to be odd. Suppose the mean value of the number of children in 100 families is 2.5, we conclude that on an average, there are 2.5 children in each family. This figure is difficult to conceive of.

Example:

Evaluate the average salary of 10 employees of the firm

$2800, $2900, $2500, $2400, $2550, $2600, $2700, $2300, $2200, $2100.

Step 1: Consider the salary for the 10 employees

X_1 = $2800, X_2 = $2900, X_3 = $2500, X_4 = $2400, X_5 = $2550, X_6 = $2600, X_7 = $2700, X_8 = $2300, X_9 = $2200, and X_{10} = $2100, here $n = 10$.

Step 2: Average salary = $\overline{X} = [1/10] \times \sum_{i=1}^{10} [X_i] = 25050/10 = \$2505$.

Hence, the average monthly salary is $2505.

Example:

Consider the data related to the monthly sales of 200 companies.

Monthly Sales (in lakhs of $)	300–350	350–400	400–450	450–500	500–550	550–600	600–650	650–700	700–750
No. of Companies	5	14	23	50	52	25	22	7	2

Evaluate the arithmetic mean.

Step 1: The given class intervals are continuous and have a uniform length. The values of X can be evaluated either directly or using difference method.

Direct Method:

Find the mid-value of the class interval (X) and create a column ($X \times f$), then

Monthly Sales (in lakhs of $)	**No. of Companies (f_i)**	X_i	$f_i \times X_i$
300–350	5	325	1625
350–400	14	375	5250
400–450	23	425	9775
450–500	50	475	23750
500–550	52	525	27300
550–600	25	575	14375
600–650	22	625	13750
650–700	7	675	4725
700–750	2	725	1450
Total	200		102000

Class = 300 – 350; mid-point = (300 + 350)/2 = 325.

For the subsequent classes add 50 to the latest value.

$$\text{Mean} = \overline{X} = \frac{\sum_{i=1}^{9}\{f_i \times X_i\}}{\sum_{i=1}^{9} f_i} = 102000/200 = \$510.$$

Hence, the average monthly sale is $510 lakhs.

Alternative Method

Using X, find the value d, with the help of the relation. $d_i = (X_i - A)/h$, where A is anyone value of X_i ($i = 1, 2, \ldots, 9$); h = length of the class interval = 50 and $n = 9$.

X_i	No. of Companies (f_i)	$d_i = (X_i - A)/h$; $A = 525$	$f_i \times d_i$
325	5	–4	–20
375	14	–3	–42
425	23	–2	–46
475	50	–1	–50
525	52	0	0
575	25	1	25
625	22	2	44
675	7	3	21
725	2	4	8
Total	200		–60

$$\overline{X} = A + h \times \frac{\sum_{i=1}^{9}\{f_i \times d_i\}}{\sum_{i=1}^{9} f_i} = 525 + 50 \times (-60/200) = \$510$$

Hence, the average monthly sale is $510.

Example:

The expenditure of 1000 families is given as follows:

Expenditure ($)	40–59	60–79	80–99	100–119	120–139
No. of Families	?	150	?	250	50

The mean of the distribution is $87.50. Calculate the missing frequencies.

Step 1: Let x and y refer to the missing frequencies. The given class intervals are not continuous and the length is uniform.

Step 2: Convert the class intervals into a continuous one.

Difference = 1; half of the difference = ½.

Subtract (1/2) and add (1/2) to the lower and upper values of the intervals. Find the midpoint of the class intervals, where $h = 20$; mid-point of the first interval = (39.5 + 59.5)/2 = 49.5 and add the length of the subsequent intervals.

Expenditure (in $)	**Number of Families (f)**	**Midpoint of the Class Interval (X)**	**$d_i = (X_i - A)/h$ $A = 89.5$**	**$f_i \times d_i$**
39.5–59.5	x	49.5	−2	$-2x$
59.5–79.5	150	69.5	−1	−150
79.5–99.5	y	89.5	0	0
99.5–119.5	250	109.5	1	250
119.5–139.5	50	129.5	2	100
Total	$450 + x + y$			$200 - 2x$

Given $\sum_{i=1}^{5}[f_i] = 1000$ families.

$$\text{Implies that: } 450 + x + y = 1000;\ x + y = 1000 - 450;\ x + y = 550 \quad (4.1)$$

by definition, $\overline{X} = A + h \times \left\{ \sum_{i=1}^{5} [f_i \times d_i] / \sum_{i=1}^{5} [f_i] \right\}$

$$\text{Mean} = 89.5 + 20 \times ((200 - 2x)/1000) \quad (4.2)$$

given the mean = 87.5.

using the value of mean in the second equation,

$$87.5 = 89.5 + 20((200 - 2x)/1000);\ 87.5 - 89.5 = 20 \times ((200 - 2x)/1000)$$

$$-2 = (200 - 2x)/50;\ -100 = 200 - 2x;\ 2x = 200 + 100 = 300;\ x = 300/2 = 150 \quad (4.3)$$

Using the value of x in the first equation, we have $150 + y = 550$;

$y = 550–150 = 400$. Hence, the missing frequencies are

Class	40–59	80–99
Missing frequency	150	400

Example:

The average monthly production of cotton piece of goods in India for the first 8 months was 409.8 million yards, and for the remaining 4 months, it was 412.1 million yards. Calculate the average monthly production for the year as a whole.

By definition,

$$\text{Average production} = \text{total production/no. of months} \quad (4.4)$$

total production = (average production) × (no. of months)

Total production for the first 8 months = 409.8 × 8 = 3278.4 million yards.

Total production for next 4 months = 412.1 × 4 = 1648.4 million yards

Total production for 12 months = (3278.4 + 1648.4) = 4926.8 million yards.

Mean production for 12 months = (4926.8/12) = 410.57 million yards.

Average production for 12 months is 410.57 million yards.

Let the total number of men and women in the group be 100.

Let n be the number of men and then $(100 - n)$ be the number of women in the group.

By definition, the mean of two composite groups is

$$\overline{X} = (n_1x_1 + n_2x_2)/(n_1 + n_2) \quad (4.5)$$

Here $\overline{X} = 30$ years, $\overline{X}_1 = 32$ years; $\overline{X_2} = 27$ years; $n_2 = 100 - n$ and $n_1 = n$.

Using all the values in equation (4.5)

$30 = (n \times 32 + (100 - n) \times 27)/100$; $32n + 2700 - 27n = 3000$; $5n = 3000 - 2700$
$= 300$; $n = 300/5 = 60$.

Using $n = 60$, we have $n_2 = 100 - 60 = 40$. The percentage of the men = 60;

The percentage of the women = 40.

Example:

The average weight of a group of 25 boys was calculated to be 78.4 lbs. It was later discovered that 1 weight was misread as 69 lbs instead of the correct value of 96 lbs. Calculate the correct average.

Given, Mean of 25 boys = 78.4 lbs. Total weight of 25 boys = 78.4 × 25 = 1960 lbs.

The value 1960 lbs include the incorrect value of 69 lbs.

Subtracting the incorrect value and adding the correct value 96 lbs, we can have the corrected total weight. Corrected total weight = 1960 – 69 + 96 = 1987.

Corrected mean = 1987/25 = 79.48 lbs. The corrected average is 79.48 lbs.

4.5 Median

The median is the size of the middle item when the items form an array. Half the total number of cases will lie below the median and half will be above. It is a 'positional' average.

4.5.1 Discrete Data

Consider the given set of n number of discrete values X_i $[i = 1, 2, \dots, n]$, then the median is defined as the mid-value of the data after setting the data in the form of ascending order, if it contains with odd number of data set. If the number of data set is even, it considers the average value of the middle two items.

4.5.2 Discrete Data with Frequency

Consider the given set of n discrete value corresponding with n different frequencies
X_i and f_i $[i = 1, 2, \dots, n]$

X	X_1	X_2	X_3	…	…	X_n
f	f_1	f_2	f_3	…	…	f_n

Construct the cumulative frequency column. Find the value of $n/2$. Select the cumulative frequency just greater than the value of $n/2$. Then the value of x corresponds to the selected cumulative frequency.

X	X_1	X_2	X_3	…	…	X_n
f	f_1	f_2	f_3	…	…	f_n
Cumulative Frequency	f_1	f_1+f_2	$f_1+f_2+f_3$			$\sum_{i=1}^{n} f_i$

4.5.3 Continuous Data with Frequency

Consider the given set of data. Verify whether the given class interval is continuous. If not continuous, change it into continuous one with proper methods. After that, verify if the length of the class intervals is uniform. If not uniform, make proper adjustment of the variable length.

Class Interval	$L_1 - U_1$	$L_2 - U_2$	$L_3 - U_3$	…	$L_n - U_n$
f	f_1	f_2	f_3	…	f_n

where L_i is the lower limit of the ith class interval and, U_i is the upper limit of the ith class interval.

For each interval, choose the mid-value and call it X_i [mid value of the ith class interval]. $X_i = [L_i + U_i]/2$

Construct the cumulative frequency column. Find the value of $n/2$. Select the cumulative frequency just greater than the value of $n/2$. Then select the interval that corresponds to the selected cumulative frequency as median class.

Class Interval	$L_1 - U_1$	$L_2 - U_2$	$L_3 - U_3$	…	$L_n - U_n$
f	f_1	f_2	f_3	…	f_n
Cumulative Frequency	f_1	f_1+f_2	$f_1+f_2+f_3$		$\sum_{i=1}^{n} f_i$

where:

l is lower limit of the median class; *h* is the length of the class interval; *n* is the total frequency; C_f is the cumulative frequency of the previous class to the median class, and

f is the frequency of the median class

The value of median can be computed using the following relationship:

$$Median\ = Me = l + h \times \left[\frac{(n/2) - C_f}{f}\right]$$

4.5.3.1 Relative Advantages

1. The median is easily calculated and can be understood easily.
2. The value of the median is not affected by the magnitude of the extreme values.

   Example:

   Median of the 5 employees' income is, \$30, \$35, \$40, \$45 and \$50 is \$40. If we change the value of the fifth item \$50 by \$100, still the median is \$40.
3. The median can be evaluated even if the data are incomplete.

   Example:

   For the preceding problem, it is possible to evaluate the median, but the arithmetic mean cannot be evaluated.
4. The median may be located when the items in a series cannot be measured quantitatively like the fairness of the skin and intelligence.

4.5.3.2 Relative Disadvantages

1. With the medians of 2 groups, the overall median cannot be evaluated.
2. If there is a high degree of variation among the data set, median cannot be viewed as a representative.

   Example:

   Median of 10, 20, 30, 100, 1000, 2000, and 3000 is 100, which is not a representative of the group.
3. It cannot be considered as a representative when there are a few items.

   Example:

   Given are the per day salary of a batch of employees in dollars. Find out the median salary \$28, \$44, \$50, \$30, \$22, \$63, \$58, \$52, \$60, \$23, \$32, \$57, \$62, \$39, \$24, \$41, \$31, \$20, \$61, \$38, \$59, \$46, \$48, \$37, \$45.

   The data type is DD. Place all the given 25 values in ascending order.

   20, 22, 23, 24, 28, 30, 31, 32, 37, 38, 39, 41, 44, 45, 46, 48, 50, 52, 57, 58, 59, 60, 61, 62, 63.

   Select the middle item. Here it is the 13th item, which is 44.

   Hence, the median is 44. The median of the per day salary is \$44.

Example:

Find the median from the following table:

X, Salary per Day ($)	22	27	32	37	42	47	52	57	62
f, Number of Employees	80	166	298	507	605	700	630	450	190

The data type is DDF. Based on the given table, construct the cumulative frequency table.

X ($)	**f**	**Cumulative Frequency**
22	80	80
27	166	246
32	298	544
37	507	1051
42	605	1656
47	700	2356
52	630	2986
57	450	3436
62	190	3626
Total	3626	

$$n/2 = 3626/2 = 1813$$

Here, the cumulative frequency just greater than 1813 is 2356.

Hence, the median is the value of X corresponding to the cumulative frequency 2356. The median is $47.

Example:

Evaluate the median value, after classifying the data given into a continuous one with class length of 10 marks:

28, 44, 50, 30, 22, 63, 58, 52, 60, 23, 32, 57, 62, 39, 24, 41, 31, 20, 61, 38,59, 46, 48, 37, 45.

Construct a continuous frequency table by taking $h = 10$.

Minimum mark = 20; maximum mark = 63.

The data type is a continuous one.

First, we have to find the class interval in which the median lies, for which find the value of $[1/2]\sum_{1}^{5}[f_i] = [1/2] \times 25 = 12.5$

Class Interval of Mark	**Tally Marks**	**Frequency (f)**	**Cumulative Frequency (C_f)**
20–30	~~////~~	5	5
30–40	~~////~~ /	6	11 [C_f]
40–50	~~////~~	5 [f]	16
50–60	~~////~~	5	21
60–70	////	4	25
	Total	25	

Cumulative frequency just greater than 12.5 is 16, which corresponds to the class (40–50) and is considered to be the median class.

$$l = 40;\ h = 10;\ C_f = 11;\ n/2 = 12.5;\ f = 5;$$

$$\text{Median} = Me = l + h \times \left[\frac{(n/2) - C_f}{f}\right] = 40 + 10 \times ((12.5 - 11)/5) = 40 + 10 \times 0.3 = 43.$$

where l is the lower limit of the median class; h is the length of the class interval; n is the total frequency; C_f is the cumulative frequency of the previous class of the median class; and f is the frequency of the median class.

Hence, the median is 43. The median value implies that 50% of the students secured below 43 marks and 50% of the student secured above 43 marks.

Example:
Consider the following data, which relates to the age distribution of 1000 workers in an industry:

Age (Years)	<25	25–30	30–35	35–40	40–45	45–50	50–55	>55	TOTAL
No. of Workers	120	125	180	160	150	140	100	25	1000

Evaluate the median age.

The data type is CDF. Here the given structure of the table is open and closed end. Convert the data into a continuous one. Moreover, the intervals are continuous and of uniform length. The median value can be evaluated directly. Construct the cumulative frequency column based on the given table.

Age (Years)	**Number of Workers**	**Cumulative Frequency**
20–25	120	120
25–30	125	245
30–35	180	425 [C_f]
35–40	160 [f]	585
40–45	150	735
45–50	140	875
50–55	100	975
55–60	25	1000
	Total	1000

Cumulative frequency just greater than 500 is 585.

Hence, the median class is (35–40).

Here $l = 35$; $n/2 = 500$; $C_f = 425$; $f = 160$ and $h = 5$,

$$\text{Median } Me = l + h \times \left[\frac{(n/2) - C_f}{f}\right] = 35 + 5 \times ((500 - 425)/160) = 35 + 5 \times (0.46875)$$

$$= 35 + 2.34375 = 37.34375 = 37.34 \text{ years}$$

The required median age is 37.34 years. The median value implies that 50% of the workers are younger than 37.34 years age and 50% of the workers are older than 37.34 years age.

Example:

An incomplete frequency distribution is given as follows:

Variable	10–20	20–30	30–40	40–50	50–60	60–70	70–80	Total
Frequency	12	30	?	65	?	25	18	229

Given that the median value is 46, find the missing frequencies using the median formula.

The given class intervals are continuous and have a uniform class length and so keep them as such. Let X and Y be the missing frequencies of the class 30–40 and 50–60, respectively. The value of the median 46 implies that obviously the median lies in the interval (40–50).

Construct the cumulative frequency column.

Given, the total frequency is 229.

$$150 + X + Y = 229;\ X + Y = 229 - 150 = 79;\ X + Y = 79 \tag{4.6}$$

The median class (40–50) implies that

$$n = 229;\ l = 40;\ C_f = 42 + X;\ f = 65;\ h = 10;\ \text{and median} = 46.$$

By definition, Median $Me = l + h \times \left[\dfrac{(n/2) - C_f}{f}\right] = 40 + 10 \times (((229/2) - (42 + X))/65)$

Variable	**Frequency**	**Cumulative Frequency**
10–20	12	12
20–30	30	42
30–40	X	$42 + X$
40–50	65	$107 + X$
50–60	Y	$107 + X + Y$
60–70	25	$132 + X + Y$
70–80	18	$150 + X + Y$
Total	229	

$$46 = 40 + 10 \times ((114.5 - 42 - X)/65);\ 6 = \{10/65\} \times (72.5 - X)$$

$$6 = \{2/13\} \times (72.5 - X);\ 6 \times 13 = 2 \times (72.5 - X)$$

$$78 = 145 - 2\,X;\ 2X = 145 - 78 = 67;\ X = 67/2 = 33.5 = 34.$$

Approximately $X = 34$.

Using the values of $X = 34$ in equation (4.6), we have

$$34 + Y = 79; Y = 79 - 34 = 45.$$

Hence, the missing frequencies are,

Variable	30–40	50–60
Frequency	34	45

4.5.3.3 Property of Median

The sum of the absolute deviations about the median is minimum.

It is denoted by $\sum_{i=1}^{n} \{ | X_i - \text{Median} | \}$ is minimum.

4.5.4 Graphical Method to Find the Median

Method 1:
Consider the continuous frequency distribution. Construct the less-than-cumulative-frequency column. Draw the less-than-cumulative-frequency curve (Ogive curve) by taking the class interval on the x-axis and the frequency on the y-axis.

Find ($n/2$) and draw a horizontal line at $Y = n/2$; it will touch the less-than-Ogive curve. Then draw a perpendicular line to the x-axis from that intersecting point. The intersecting point on the x-axis is the required median.

Example:

Evaluate median using graphical method.

Size	0–10	10–20	20–30	30–40	40–50	50–60	60–70
Frequency	5	7	19	18	16	10	5

The given class intervals are continuous and have uniform length. So, keep it as such. Construct the cumulative frequency column and mid-class interval column.

Consider the upper limits of the class intervals on the x-axis and frequency along the y-axis. $n/2 = 80/2 = 40$; draw a horizontal line at $y = 40$. It is clear that, it intersects the cumulative frequency curve. Draw a perpendicular line from the point of intersection to the x-axis. The point at which it intersects on the x-axis is the required median.

Approximately the median value is 35 (Figure 4.1).

Size	Frequency	Less Than	Less-than-Cumulative Frequency
0–10	5	0	0
10–20	7	10	5
20–30	19	20	12
30–40	18	30	31
40–50	16	40	49
50–60	10	50	65
60–70	5	60	75
		70	80

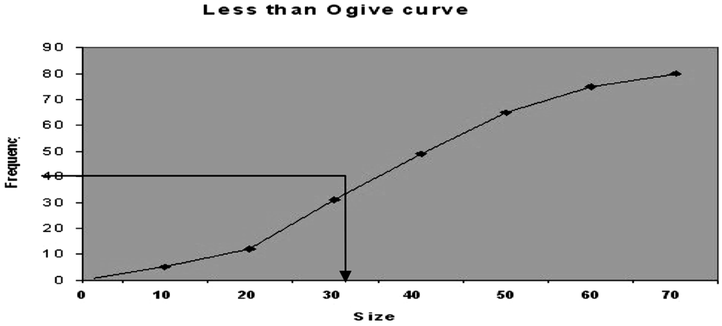


FIGURE 4.1
Less-than Ogive curve.

Method 2:
Consider the continuous frequency distribution. Construct the less-than-and more-than-cumulative-frequency columns. Draw the less-than and more-than-Ogive curves by taking frequency along the y-axis and the class intervals along the x-axis. Draw the perpendicular line from the point of intersection of both the Ogive curves to the x-axis. The point at which it meets the x-axis is the required median. Consider the previous example. Already the less-than-cumulative frequency is evaluated. Construct the more-than-cumulative-frequency column.

Size	Frequency	More Than	More-than-Cumulative Frequency
0–10	5	0	80
10–20	7	10	75
20–30	19	20	68
30–40	18	30	49
40–50	16	40	31
50–60	10	50	15
60–70	5	60	5
		70	0

Draw both the less-than-and more-than-Ogive curve (LOC and MOC).
The approximate value of median is 35 (Figure 4.2).

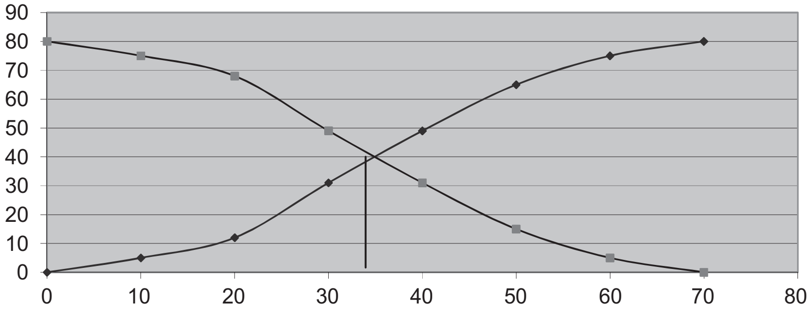


FIGURE 4.2
Less-than and more-than Ogive curves.

4.6 Quartiles, Deciles and Percentiles

Quartiles are measured like the median. Median divides the series into 2 equal parts. Extending this concept of median, we have quartiles. Quartiles can be classified into Q_1, Q_2, and Q_3. This divided the series into four equal parts. Among these three, Q_2 is the median.

Similarly, the distribution can be divided into 10 equal parts called 'deciles' and into 100 equal parts called 'percentiles'.

Nature of items	$i = 1, 2, 3$; ith quartile item	ith decile item $i = 1, 2, \dots, 9$	ith percentile item $i = 1, 2, 3, \dots, 99$
Discrete items	$((n + 1)/4) \times i$	$((n + 1)/10) \times i$	$((n + 1)/100) \times i$
Continuous items	$(n/4) \times i$	$(n/10) \times i$	$(n/100) \times i$

The evaluation procedures of quartiles, deciles, and percentiles are similar to the evaluation of the median. All the these statistics have the advantages and disadvantages similar to the median. An Ogive curve can be used to locate them. Quartiles are particularly useful in statistics for calculating quartile deviation and measuring the skewness of a distribution.

Example:

Find the quartiles of the following data: 68, 50, 32, 21, 54, 38, 59, 66, 44

Rewrite the given data in the ascending order.

21, 32, 38, 44, 50, 54, 59, 66, 68; Here $N = 9$.

Q_1: find $(N + 1)/4 = (9 + 1)/4 = 2.5$ item

Average value of 2nd and 3rd item

$Q_1 = [½] (32 + 38) = 35$; $Q_1 = 35$.

Q_2: find $N + 1/2 = (9 + 1)/2 = 10/2 = 5$; $Q_2 = $ 5th item $= 50$

Q_3: find $3(N + 1)/4 = 30/4 = 7.5$th item

$Q_3 = $ average of 7th and 8th item.

$Q_3 = [1/2]\ (66 + 59) = ½(125) = 62.5 = 62.5$

Hence, the value of quartiles are $Q_1 = 35$; $Q_2 = 50$ and $Q_3 = 62.5$.

Example:

Find the quartiles of the following distribution:

Values (*X*)	10	15	20	25	30	35	40
Frequency (*f*)	5	10	25	30	20	15	2

Construct the cumulative frequency column and find the value of *N*/4, *N*/2, and (3/4) × *N*.

X	*f*	**Cumulative Frequency**
10	5	5
15	10	15
20	25	40
25	30	70
30	20	90
35	15	105
40	2	107
Total	107	

Here $N = 107$.

Q_1: $N/4 = 107/4 = 26.75$

The value of cumulative frequency just greater than 26.75 is 40.

$Q_1 = $ The value of x corresponds to the cumulative frequency 40 is 20.

$Q_1 = 20$; Q_2: $N/2 = 107/2 = 53.5$.

The value of cumulative frequency just greater than 53.5 is 70.

$Q_2 = $ the value of x corresponds to the cumulative frequency 70 is 25.

$Q_2 = 25$; Q_3: $(3/4) \times N = 3 \times 107/4 = 80.25$

The value of cumulative frequency just greater than 80.29 is 90.

$Q_3 = $ the value of x corresponds to the cumulative frequency 90 is 30.

$Q_3 = 30$. Hence, the quartiles are $Q_1 = 20$; $Q_2 = 25$, and $Q_3 = 30$.

Example:

Calculate the quartiles and the D3 for the following data.

Difference (Years)	0–5	5–10	10–15	15–20	20–25	25–30	30–35	35–40
Frequency	449	705	507	281	109	52	16	4

The given data set is continuous with uniform length. Construct the cumulative frequency column. Also find $N/4$, $N/2$, $3 \times [N/4]$, and $3 \times [N/10]$.

Difference (Years)	Frequency	Cumulative Frequency
0–5	449	449
5–10	705	1154
10–15	507	1661
15–20	281	1942
20–25	109	2051
25–30	52	2103
30–35	16	2119
35–40	4	2123
Total	2123	

Evaluation of Q_1

$$N/4 = 530.75$$

The cumulative frequency just greater than 530.75 is 1154.

The corresponding first quartile class is 5–10.

Here $l = 5$; $h = 5$; $f = 705$; $C_f = 449$.

$$\begin{aligned} Q_1 &= l + h \times ((N/4 - C_f) \ / \ f) = 5 + 5 \times ((530.75 - 449)/705) = 5 + 5 \times (0.1160) \\ &= 5 + 0.58; \\ Q_1 &= 5.58 \end{aligned}$$

Evaluation of Q_2

$$N/2 = 2123/2 = 1061.5$$

The cumulative frequency just greater than 1061.5 is 1154.

The corresponding second quartile class is 5–10. Here $l = 5$; $h = 5$; $f = 705$; $C_f = 449$.

$$\begin{aligned} Q_2 &= l + h \times ((N/2 - C_f)/f) = 5 + 5 \times ((1061.5 - 449)/705) = 5 + 5 \times (0.8688) \\ &= 5 + 4.344 = 9.344 \end{aligned}$$

Evaluation of Q_3

$$3N/4 = (3 \times 2123)/4 = 1592.25$$

The cumulative frequency just greater than 1592.25 is 1661.

The corresponding third quartile class is 10–15. Here $l = 10$; $h = 5$; $f = 507$; $C_f = 1154$

$$Q_3 = 10 + 5 \times ((1592.25 - 1154)/507) = 10 + 5 \times (0.8644) = 14.322.$$

Evaluation of D_3

$$3N/10 = (3 \times 2123)/10 = 636.9$$

The cumulative frequency just greater than 636.9 is 1154.

The corresponding third deciles class is 5–10. Here $l = 5$; $h = 5$; $f = 705$; $C_f = 449$

$$D_3 = l + h \times ((3N/10 - C_f)/f) = 10 + 5 \times ((636.9 - 449)/705)$$
$$= 10 + 5 \times (0.2665) = 10 + 1.333$$

$$D_3 = 11.333$$

Hence, the required quartiles and deciles are $Q_1 = 5.58$; $Q_2 = 9.344$; $Q_3 = 14.322$; $D_3 = 11.333$.

4.7 Mode

The value of the variable that occurs most frequently called ‘mode’. It is the position of greatest density, the predominant or most common value. It is also a positional average.

4.7.1 Discrete Data

Consider the given set of n number of discrete values X_i [$i = 1, 2, \dots, n$]. For the discrete, series mode is not well defined. The approximate value of the mode can be computed using the following relationship: Mode = 3 × Median – 2 × Mean

4.7.2 Discrete Data with Frequency

Consider the given set of n discrete value that corresponds with n different frequencies X_i and f_i [$i = 1, 2, \dots, n$];

X	X_1	X_2	X_3	…	…	X_n
f	f_1	f_2	f_3	…	…	f_n

Select the maximum frequency. The value of X corresponding to the maximum frequency is considered to be the mode of the data set.

4.7.3 Continuous Data with Frequency

Consider the given set of data. Verify whether the given class interval is continuous or not. If not continuous, change it into continuous one with proper methods. After that, verify whether the length of the class intervals is uniform or not. If not uniform, do proper adjustment of the variable length.

Class Interval	L_1–U_1	L_2–U_2	L_3–U_3	…	L_n–U_n
f	f_1	f_2	f_3	…	f_n

where L_i is the lower limit of the ith class interval, and U_i is the upper limit of the ith class interval

For each interval, chose the mid-value, and call it as X_i [mid-value of the ith class interval]. $X_i = [L_i + U_i]/2$

Select the maximum frequency. The class interval that corresponds to the maximum frequency is considered to be the modal class of the data set.
The value of mode can be computed using the following relationship:

$$\text{Mode} = 1 + h \times \left[\frac{(f_0 - f_1)}{(2 \times f_0 - f_1 - f_2)} \right]$$

where *l* is the lower limit of the modal class; *h* is the length of the class interval; f_0 is the frequency of the modal class; f_1 is the frequency of the class preceding to the modal class; and f_2 is the frequency of the class succeeding to the modal class.
The relative advantages of mode:

- It is useful in the study of popular sizes.
- It is simple.
- It is not affected by the extreme values and can be calculated even if extreme values are unknown.

Example:

A banker can use mode instead of mean to decide the average balances of deposits.
The relative disadvantages of mode:

- It is not well defined. Sometimes it is not possible to locate it properly.
- A distribution may be bimodal or multimodal.
- It is not suitable for mathematical treatment.

Example:

Calculate mode from the following data x = 2, 3, 6, 5, 3, 2, 3, 9, 11, 23, 3.

	Tally Mark	Frequency (*f*)
2	//	2
3	////	4
5	/	1
6	/	1
9	/	1
11	/	1
23	/	1

Construct the discrete data with frequency.

Select the maximum frequency. Here it is 4. This corresponds to the value 3.

Hence, the mode is 3. $M_o = 3$.

Example:

Find the mode for the data given below by using the method of grouping.

Size	1	2	3	4	5	6	7	8	9	10	11	12	13
Frequency	52	56	60	63	57	55	50	52	41	57	63	52	48

The given data contain 2 modes (i.e., bimodal distribution) because of the sizes 4 and 11 having the highest frequency 63. So, both can be considered as a mode. We want to know which is more representative of the distribution. By considering size and frequency

- Create column 3 by combining the subsequent 2 frequencies starting from the first one.
- Create column 4 by combining the subsequent 2 frequencies starting from the second one.
- Create column 5 by combining the subsequent 3 frequencies by adding from the first one.
- Create column 6 by combining the subsequent 3 frequencies by adding from the second one.
- Create column 7 by combining subsequent 3 frequencies by adding from the third one.

Size	**Frequency**	**Cl 3**	**Cl 4**	**Cl 5**	**Cl 6**	**Cl 7**
1	52					
2	56	108				
3	60		116	168		
4	63	123			179	
5	57		120			180
6	55	112		175		
7	50		105		162	
8	52	102				157
9	41		93	143		
10	57	98			150	
11	63		120			161
12	52	115		172		
13	48		100		163	
	Maximum	123	120	175	179	180
	Size corresponding to the maximum value is	3,4	4,5,10,11	4,5,6	2,3,4	3,4,5

The size 4 repeats a maximum number of groups. This implies that the modal size is 4.

NOTE: This method can be applied to the irregular distribution also.

Consider the irregular distribution.

Size (x)	1	2	3	4	5	6
Frequency (f)	3	8	15	40	20	30

Example:

The distribution of the number of months of holding of shares of a company by its shareholders is presented. Find the mean, median, and mode of the number of holdings of the shares by its shareholder:

For the given table, a structure is open- and closed-ended, so changes it into a continuous one. Let us take the initial class as 0–2 and the last class as 18–20. Create a new column with midpoint of the class intervals and cumulative frequency column. Here $h = 2$.

No. of Months of Holding	**No. of Share Holders ('000)**	**No. of Months of Holding**	**No. of Share Holders (000)**
Less than 2	5	10–12	6
2–4	8	12–14	20
4–6	10	14–16	12
6–8	7	16–18	9
8–10	9	More than 18	2

$$\text{Mean} = \sum_{i=1}^{10}[X_i \times f_i] / \sum_{i=1}^{10}[f_i] = 906/88 = 10.295 = 10.3$$

$$n = [1/2] \times \sum_{i=1}^{10}[f_i] = 88/2 = 44.$$

Cumulative frequency just >44 is 45.

The median class is (10–12). Here $l = 10$; $h = 2$; $C_f = 39$; $f = 6$

$$\text{Median} = 1 + h \times \left[\frac{(n/2) - C_f}{f}\right]$$

$$= 10 + 2 \times ((44 - 39)/6) = 10 + 2 \times [5/6] = 10 + 1.67 = 11.67$$

No. of Months of Holding	No. of Shareholders (000)(f)	Midpoint (X) of Class Intervals	$f \times X$	Cumulative Frequency
0–2	5	1	5	5
2–4	8	3	24	13
4–6	10	5	50	23
6–8	7	7	49	30
8–10	9	9	81	39
10–12	6	11	66	45
12–14	20	13	260	65
14–16	12	15	180	77
16–18	9	17	153	86
18–20	2	19	38	88
Total	88		906	

The maximum frequency is 20, which correspond to the class (12–14).

The modal class is 12–14. Here $l = 12$; $h = 2$; $f_0 = 20$; $f_1 = 6$; $f_2 = 12$.

$$\text{Mode} = l + h \times \left[\frac{(f_0 - f_1)}{(2f_0 - f_1 - f_2)}\right] = 12 + 2 \times ((20 - 6)/(2 \times 20 - 6 - 12)) = 12 + 2 \times (14/22)$$

$$= 12 + 2 \times (7/11) = 12 + 1.27 = 13.27;$$

Mode = 13.27.

Hence, 10.3 months, 11.67 months, and 13.27 months stand for mean, median, and mode, respectively, of the number of months of holding of the shares by its shareholders.

Example:

The expenditure of 100 families is given:

Expenditure ($)	0–10	10–20	20–30	30–40	40–50
No. of families	14	?	27	?	15

The mode of the distribution is 24. Calculate the missing frequencies. The given class intervals are continuous and are in uniform length. Let x and y be the unknown frequencies of classes 10–20 and 30–40, respectively. Here $h = 10$; mode = 24; and $n = 100$. The modal class is 20–30.

Expenditure	No. of Families
0–10	14
10–20	x [f_1]
20–30	27 [f_0]
30–40	y [f_2]
40–50	15
Total	$56 + x + y$

$n = 100$, implies that $100 = 56 + x + y$, that is $x + y = 100 - 56 = 44$

Then $x + y = 44$ (4.6)

By definition,

$$\text{Mode} = l + h \times \left[\frac{(f_0 - f_1)}{(2 \times f_0 - f_1 - f_2)}\right] \tag{4.7}$$

Here $f_0 = 27; f_1 = x; f_2 = y$. Using the values in Equation (4.7),

$$24 = 20 + 10((27 - x)/(54 - x - y)); 24 - 20 = 10((27 - x)/(54 - x - y))$$

$$4 = 10\ ((27 - x)/(54 - x - y)); 2 = 5((27 - x)/(54 - x - y)); 2(54 - x - y) = 5(27 - x)$$

$$108 - 2x - 2y = 135 - 5x; -2x - 2y + 5x = 135 - 108; 3x - 2y = 27 \tag{4.8}$$

$$(4.6) \times 2 + (4.8); \text{ implies that}$$

$$[2x + 2y] + [3x - 2y] = [88 - 27]; 5x = 115$$

$$x = 115/5 = 23; x = 23; \text{ using the value of } x \text{ in (4.6); we have}$$

$$23 + y = 44; \text{ implies that } y = 44 - 23 = 21; y = 21.$$

Hence, the missing frequencies are

Expenditure	Number of Families
10–20	23
30–40	21

Example:

A welfare organization introduced an education scholarship scheme for the school-going children of a village. The rates of scholarships were fixed as given:

Age Groups (Years)	Amount of Scholarship per Month ($)
5–7	30
8–10	40
11–13	50
14–16	60
17–19	70

The ages (years) of 30 school going-children are noted as 11, 8, 10, 5, 7, 12, 7, 17, 5, 13, 9, 8, 10, 15, 7, 12, 6, 7, 8, 11, 14, 18, 6, 13, 9, 10, 6, 15, 13, 5 years, respectively. Calculate mean of monthly scholarship. Find out a total monthly scholarship amount being paid to the students.

Construct the frequency distribution based on the given data.

$$\sum_{i=1}^{5}[X_i \times f_i]/\sum_{i=1}^{5}[f_i]=1290/30$$

Mean = 43. The average monthly scholarship is 43.

Age Groups (Years)	**Tally Marks**	**No. of Children (*f*)**	**Amount of Scholarship per Month ($) (*x*)**	$x \times f$
5–7	~~////~~ ~~////~~	10	30	300
8–10	~~////~~///	8	40	320
11–13	~~////~~//	7	50	350
14–16	///	3	60	180
17–19	//	2	70	140
Total		30		1290

Total monthly scholarship paid to the students is $ 1290.

Example:

Calculate mean, median and mode from the following data:

Value	**Frequency**	**Value**	**Frequency**
Below 10	2	Below 60	303
Below 20	21	Below 70	333
Below 30	56	Below 80	351
Below 40	133	Below 90	358
Below 50	233	Below 100	350

The given table is of the form open-ended. First, convert it with uniform class intervals.

Let $h = 10$. Find the mid-values of the class intervals. Select the values of A and h.

$$d = (x - A)/h;\ A = 55;\ \text{and}\ h = 10$$

Value	**Frequency (*f*)**	**Class-Intervals**	**Mid-(*x*)**	**d**	$f \times d$	**Cumulative Frequency**
<10	2	0–10	5	−5	−10	2
<20	21	10–20	15	−4	−84	23
<30	56	20–30	25	−3	−168	79
<40	133	30–40	35	−2	−266	212
<50	233	40–50	45	−1	−233	445
<60	303	50–60	55	0	0	748
<70	333	60–70	65	1	333	1081
<80	351	70–80	75	2	702	1432
<90	358	80–90	85	3	1074	1791
<100	350	90–100	95	4	1400	2140
	2140				2748	

$$\overline{X} = A + h \times \{\sum_{i=1}^{10}[f_i \times d_i] / \sum_{i=1}^{10}[f_i]\}$$

Mean = 55 + 10 × (2748/2140) = 55 + 10 × 1.284 = 55 + 12.84 = 67.84. Mean = 67.84

Median:

$$n = 2140,\ n/2 = 2140/2 = 1070.$$

Cumulative frequency, just greater than 1070 is 1081.

Hence, the median class is (60–70). Here, $l = 60$; $h = 10$; $C_f = 748$; $f = 333$.

$$Me = 1 + h \times \left[\frac{(n/2) - C_f}{f}\right] = 60 + 10 \times ((1070 - 748)/333) = 60 + 10 \times 0.967 = 69.67$$

Median = 69.67

Mode:

The maximum frequency corresponds to 358. The modal class is (80–90).

Here, $l = 80$; $f_0 = 358$; $f_1 = 351$; $f_2 = 350$; and $h = 10$

$$\text{Mode} = 1 + h \times \left[\frac{(f_0 - f_1)}{(2 \times f_0 - f_1 - f_2)}\right] = 80 + 10 \times ((358 - 351)/(2 \times 358 - 351 - 350))$$

$$= 80 + 10 \times (7/15) = 80 + 10 \times 0.4667 = 84.67;$$

Mode = 84.67

4.7.4 A Graphical Method to Evaluate the Mode

Graphical method can be used to evaluate the mode if and only if the given data set follows a continuous distribution with uniform class length.

Example:

Evaluate mode using graphical method.

Monthly Salary	2000–2100	2100–2200	2200–2300	2300–2400	2400–2500	2500–2600	2600–2700
No. of Employees	15	25	28	42	30	20	10

Because the given data is continuous distribution and the lengths of the class intervals are uniform, draw the histogram (Figure 4.3).

The approximate value of the mode is 2353.

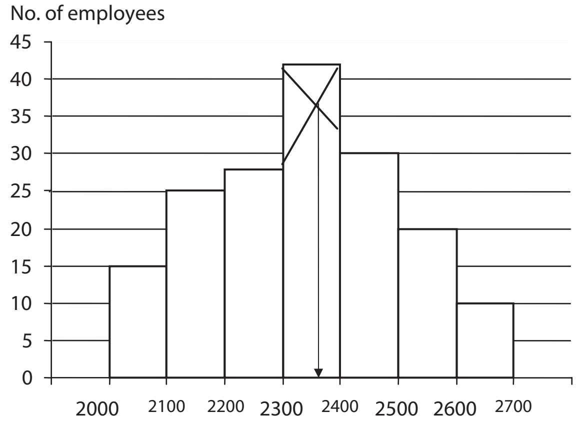


FIGURE 4.3
Histogram.

4.8 Comparison of Mean, Median, and Mode

The determination of which average is exactly suited for a specific variable depends on many factors. Certainly it depends on the data level. The following table summarizes the valid averages for each level of data:

Data Level	Averages Can Be Evaluated
Nominal	Mode
Ordinal	Mode and median
Interval	Mode, median, and mean
Ratio	Mode, median, and mean

It is often convenient to talk about the shape of the distribution. A symmetrical distribution with one mode is commonly called a 'bell-shaped curve'. For a symmetrical distribution, all the three measures (i.e., mean, median, and mode) are exactly equal in value.

$$\text{Mean} = \text{Median} = \text{Mode}$$

If a distribution is not symmetrical, then it is 'asymmetrical' or a 'skewed distribution'. If the distribution is moderately asymmetrical, the following relationship holds good approximately.

$$\text{Mode} = 3 \times \text{Median} - 2 \times \text{Mean (or) Mean} - \text{Mode} = 3 \times (\text{Mean} - \text{Median})$$

This relation is called an 'empirical relation'. Using the empirical relation, if any two measures are known, then the third one can be evaluated approximately.

Example:

Calculate the arithmetic mean and median of the following distribution. Also, find the mode value using the empirical relationship.

Value	Frequency	Value	Frequency
>50	250	>300	40
>100	240	>350	25
>150	210	>400	15
>200	170	>450	5
>250	100	>500	0

Consider the given open-ended intervals and convert them into continuous class intervals with uniform length.

Value	Frequency	Value	Frequency
50–100	10	300–350	15
100–150	30	350–400	10
150–200	40	400–450	10
200–250	70	450–500	5
250–300	60		

Evaluate the mean and median:

The mean value is 236, and the median value is 232.14.

The value of mode can be obtained using the empirical relation.

Mode = 3 × Median – 2 × Mean; Mode = (3 × 232.14) – (2 × 236) = 224.42.

Hence, the approximate mode value is 224.42.

4.9 Weighted Arithmetic Mean

The weighted arithmetic mean is the average evaluated after applying weights to the item as judged by their relative importance. It becomes essential whenever the items in the group are not exactly homogeneous.

Example:

Assume that grades representing a semester of work contain one final examination and two 1-hour examinations.

Examination	Marks Secured (X)	Weight Attached (w)	$w \times X$
Final	75	2	150
First 1-hour	80	1	80
Second 1-hour	90	1	90
	Total	4	320

$$\text{Weighted Mean} = \sum_{i=1}^{3}[X_i \times w_i] / \sum_{i=1}^{3}[w_i] = 320/4 = 80 \text{ marks.}$$

The weights attached should strictly reflect the relative importance of the items and should be rounded up for early evaluation. When the average is weighted, the importance of all the items is taken into account. The ordinary average becomes a special case of a weighted average. If we take the weight uniformly as one, the weighted mean becomes ordinary mean.

4.9.1 Advantages of the Weighted Mean

The weighted mean is used in the following instances.

- It is used in constructing index numbers. The relative weights of the expenditure like food, clothing, housing, and such are obtained by surveys and the cost-of-living index is calculated with those weights.
- In educational institutions, it is used to assess the real merit of the student.
- It is used in evaluating standardized death rates.

4.10 Geometric Mean

The geometric mean (GM) is the *n*th root of the product of *n* items.

$$\text{GM} = (X_1 \times X_2 \times \ldots\ldots\ldots\ldots \times X_n)^{1/n}$$

It is a mathematical average and not a positional average. It takes all the given values in its evaluation. It gives less weight to the end values than the arithmetic mean. Usually this value will be less than the mean. If any one of the element takes the value 0, GM = 0, and if any element is negative, the value of GM is imaginary. So, it is useful only in certain special situations.

NOTE: GM can be used in the following situations:

- to calculate the rates of change.
- certain cases of averaging ratios and percentages.
- problems involving rates of interest of invested money.
- can be used to interpolate between items that have a uniform rate of change.
- Used in the evaluation of index numbers.

Example:

A sum of money was invested for 5 years. The average rates of return for the investment for the 5 successive years were as follows:

5%, 4%, 5%, 6%, 3%

What was the average rate of interest for these 5 years?

If we assume that the amount invested as $100. The total amount earned in the 5 years is 105, 104, 105, 106, and 103. (i.e., 5% = 100 + 5 = 105). Here $n = 5$.

GM = $(105 \times 104 \times 105 \times 106 \times 103)^{1/5} = 104.5950 = 104.6$

This implies that the average rate of return is (104.6 – 100) = 4.6% per year.

Example:

Find the geometric mean of the data given.

Items in Cost of Living	Price Relative	Weight
Food	128.8	60
Clothing	175.6	20
House rent	110.0	10
Miscellaneous	210.0	10
Total		100

$$\text{GM} = (x_1^{f1} \times x_2^{f2} \times x_3^{f3} \times \ldots \times x_m^{fm})^{1/n}$$

$$\text{Here } n = 100;\ \text{GM} = (128.8^{60} \times 175.6^{20} \times 110^{10} \times 210^{10})^{1/100}.$$

$$= 128.8^{0.6} \times 175.6^{0.2} \times 110^{0.1} \times 210^{0.1} = 141.65.$$

Hence, the required geometric mean is 141.65.

4.11 Harmonic Mean

Harmonic mean (HM) is also a mathematical average. It is the reciprocal of the average of the reciprocal of the values.

$$HM = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{Xi}\right)}$$

Advantages:

- It is used in the averaging of time rates and in manipulation of price data.
- It is capable of algebraic manipulation.
- It is of lower value than geometric and arithmetic means.

Disadvantages:

- It is difficult to calculate.
- It is not easy to understand.
- When one of the items is 0, it becomes indeterminant.

Example:

A teacher finds that 3 students *X*, *Y*, and *Z* take 6, 3, and 8 minutes, respectively, to solve a problem. Compute the average rate of solving the problem.

Given

Time for solving the problem by $X = 6$ minutes.

Time for solving the problem by $Y = 3$ minutes.

Time for solving the problem by $Z = 8$ minutes.

Here $n = 3$;

$$\text{Then, HM} = 3/(1/6 + 1/3 + 1/8) = 3/(15/24) = 24/5 = 4.8 \text{ minutes.}$$

Exercise 4

1. Erika operates a website devoted to providing information and support for persons who are interested in organic gardening. According to the hit counter that records daily visitors to her site, the numbers or visits during the past 20 days have been as follows:

   65, 36, 52, 70, 37, 55, 63, 59, 68, 56, 65, 63, 43, 46, 73, 41, 47, 75, 75, and 54.

   Determine all the statistical measures.

2. For 1986–2006, the net rate of investment income for US life insurance companies was as follows. What was the mean net rate of investment income for these periods also evaluate the median value?

2006	05	04	03	02	2000	1999	98	97	96
6.95	7.35	7.25	7.41	7.14	7.52	8.08	8.63	8.89	9.1
1995	94	93	92	91	90	89	88	87	86
9.03	9.1	9.35	9.63	9.45	8.96	8.91	8.57	8.02	7.73

3. For the following data: Which price has more variability?

Month	Jan	Feb	Mar	Apr	May	Jun	Jul
Gold ($)	329.39	329.01	341.91	366.72	371.89	392.4	378.46
Slab Zinc ($)	0.5090	0.4726	0.4811	0.4722	0.4481	0.4508	0.4287

Month	Aug	Sep	Oct	Nov	Dec
Gold ($)	354.85	364.18	373.49	383.69	387.02
Slab Zinc ($)	0.4242	0.4388	0.443	0.4644	0.4776

4. Eighty packages have been randomly selected from a frozen food warehouses, and the age in weeks of each package is identified. Given the frequency distribution shown, determine the mean, median, and mode for the ages of the packages in the warehouse inventory.

Age (Weeks)	Number of Packages
0–under 10	25
10–under 20	17
20–under 30	15
30–under 40	9
40–under 50	10
50–under 60	4

5. Sport Way Manufacturing has been experimenting with new materials to use for golf ball covers. Two recently developed compounds have been shown to be equally resistant to cutting, and the development lab is now looking at the distance the balls will travel during a simulated drive. However, both distance and consistency are important for a golf ball. A sample of 10 balls with each type of cover was selected and the following distances were measured in yards using a mechanical driver that struck each ball with the same force.

Type A		Type B	
298	291	290	310
296	299	300	305
289	285	297	315
291	292	301	286
287	290	302	321

Evaluate mean, and median.

(Convert the given data into a continuous type 280–285; … ; 315–320 and then evaluate all the measures.)

6. Find the mean, median, and mode for the following data set referring to a frequency distribution for the usage times of 50 automated teller machine (ATM) customers:

Time (s)	Frequency
20–25	1
25–30	7
30–35	10
35–40	9
40–45	9
45–50	6
50–55	5
55–60	3

7. Eighty packages have been randomly selected from a frozen food warehouse, and the age in weeks of each package is identified. Given the frequency distribution shown, determine the median and mode of the ages of the packages in the warehouse inventory.

Age (Weeks)	Number of Packages
0–under 10	25
10–under 20	17
20–under 30	15
30–under 40	9
40–under 50	10
50–under 60	4

Also find the coefficient of variation (CV).

8. Lives of 2 models of refrigerators turned in for new models in a recent survey are:

Life (Years)	No. of Refrigerators: A	No. of Refrigerators: B
0–2	5	2
2–4	16	7
4–6	13	12
6–8	7	19
8–10	5	9
10–12	4	1

What is the average life of each model of these refrigerators?

9. Ed Grant is the director of the Student Financial Aid Office at Wilderness College. He has used available data on the summer earnings of all students who have applied to his office for financial aid to develop the following frequency distribution:

Summer Earnings ($)	Number of Students
0–499	231
500–999	304
1000–1499	400
1500–1999	296
2000–2499	123
2500–2999	68
3000 or more	23

Find the values of mean, mode, standard deviation, and CV.

NOTE: In this problem the last interval is open-ended, first convert it into closed-ended. Assume it as 3000–3499. Then convert the intervals into a continuous one. Then evaluate the measures.

10. The distribution of wages is given in two factories X and Y.

Wages ($)	No. of Workers X	No. of Workers Y
50–100	2	6
100–150	9	11
150–200	29	18
200–250	54	32
250–300	11	27
300–350	5	11

Identify which factory is consistent in paying the wages.

11. A sports psychologist, studying the effect of jogging on college students' grades, collected data from a group of college joggers. Along with some other variables, he recorded the average number of miles run per day. He compiled his results into the following distribution:

Miles per Day	Frequency
1–1.39	32
1.40–1.79	43
1.8–2.19	81
2.2–2.59	122
2.6–2.99	131
3–3.39	130
3.4–3.79	111
3.8–4.19	95
4.2–4.59	82
4.6–4.99	47
5 and up	53

Evaluate all the measures. (Note: Similar to problem 9.)

12. BMT, Inc., manufactures performance equipment for cars used in various types of racing. It has gathered the following information on the number of models of engines in different size categories used in the racing market it serves:

Class	Frequency
(Engine size, in^3)	(Number of Models)
101–150	1
151–200	7
201–250	7
251–300	8
301–350	17
351–400	16
401–450	15
451–500	7

Evaluate all the measures.

13. Frequency distribution showing the number of motorists in each speed category on a stretch of interstate highway is as follows. The distribution of data is as follows:

Speed (mph)	Number of Motorists
45–under 50	1
50–under 55	9
55–under 60	14
60–under 65	23
65–under 70	16
70–under 75	16
75–under 80	12
80–under 85	8
85–under 90	1

Find all the measures.

14. A business group is supporting the addition of a light-rail shuttle in the central business district and has 2 competing bids with different numbers of seats in each car. They arrange a fact-finding trip to Denver, and in a meeting they are given the following frequency distribution of the number of passengers per car:

Number of Passengers	Frequency
1–10	20
11–20	18
21–30	11
31–40	8
41–50	3
51–60	1

a. One bid proposes light-rail cars with 30 seats and 10 standees. What percentage of the total observations is more than 30 and less than 41 passengers?

b. The business group members have been told that street cars with fewer than 11 passengers are uneconomical to operate and with more than 30 passengers lead to poor customer satisfaction. What proportion of trips would be economical and satisfying?

c. Evaluate the mean, median, and mode for the data.

15. For the data given, compute the value of mean.

Monthly Wages ($)	No. of Workers
Below 850	12
850–900	16
900–950	39
950–1000	56
1000–1050	62
1050–1100	75
1100–1150	30
1150 and above	10

(Hint: convert the first and the last intervals as 800–850 and 1150–1200.)

16. A service station recorded the following frequency distribution for the number of gallons of gasoline sold per car in a sample of 680 cars

Gasoline	0–4	5–9	10–14	15–19	20–24	25–29
Frequency	74	192	280	105	23	6

Evaluate all measures.

17. Automobile traveling on a road that has a posted speed limit of 55 miles per hour is checked for speed by a state police radar system. Following is a frequency distribution of speeds.

Speed Miles per Hour	45–50	50–55	55–60	60–65	65–70	70–75	75–80
Frequency	10	40	150	175	75	15	10

Compute all the measures.

5

Dispersion

5.1 Introduction

An average gives a single value that is considered to be representative of the whole group, but it is not going to reveal how the items in the groups are scattered or spread.

Example:

Six boys have the marks 100,100,100,0,0,0 and another group of six boys have the marks 50, 60, 50, 50, 40, 50. Both have 50 marks as average. Comparing both the groups, group one's marks are much scattered and in group two not that much scattered. This scatter is referred to as dispersion.

The measure of dispersion is used to evaluate to what extent the data varies from the mean or from any other statistic. They are the range, the mean deviation, the quartile deviation and the standard deviation. To evaluate the relative amounts of dispersion, we use the coefficient of mean, quartile and standard deviation.

5.2 Range

The simplest measure of dispersion is the range of data, that is, the distance between the smallest and the largest values of the distribution. It is an absolute measure of dispersion.

Advantages

- It is very easy to calculate.
- It is easy to understand.

Disadvantages

- It is affected by the extreme values.

Example:

A production manager might say that the average daily wage in a machinery department is $75 and an individual's daily wage ranges from $50 to $100. This gives a rough measure of scatter that can be compared with other departments. Likewise, in the packaging department the mean wage rate may be $65, with daily wages ranging from $60 to $80. Comparing these two departments, the average of the packing department is probably more representative of the wage distribution than the average of the machinery department since, there is less scatter in the wages received by the packaging department. Moreover, the range is insensitive to the behaviour of the values between the extremes.

No.	Data	Range
1.	30,40,40,60	30
2.	30,60,60,60	30
3.	30,50,50,60	30

In all the three cases the value of the range is the same. So, it doesn't give a correct idea of dispersion. The range varies too much from sample to sample taken from the same population at random and hence it is less reliable than the other measures of deviations.

But it is very much useful when the sample size is very small.

- It is used in the quality control tests.
- It is used to measure the variations in temperature during a day/a year.
- It is used in stock exchange quotations.

5.3 Quartile Deviation (QD)

It is the measure based on the quartiles. It is the average of the difference between the third quartile and the first quartile. Mathematically it can be expressed as QD = $(Q_3 - Q_1)/2$.

Advantages

- It is also called the semi-inter-quartile range.
- It is easy to compute.
- It is not affected by the extreme values.
- It is used when the extreme values are thought to be unrepresentative.

Disadvantages

- It is not suited for further algebraic treatment.

NOTE: If the QDs of two samples are known, we cannot calculate the QD for the combined sample. That is, composite QD cannot be evaluated.

5.4 Coefficient of Quartile Deviation

The coefficient of quartile deviation (CQD) is a relative measure of dispersion based on the quartile deviation. It is defined as,

$$\text{CQD} = (Q_3 - Q_1)/(Q_3 + Q_1)$$

Moreover, it is a positional measure and is preferred when the distribution is badly skewed.

Example:

Determine the QD and CQD for the following data:

Monthly Wages ($)	Below 850	850–900	900–950	950–1000	1000–1050	1050–1100	1100–1150	>=1150
No. of Workers	12	16	39	56	62	75	30	10

Consider the open-ended table and convert the same into a continuous distribution having uniform length. Find the cumulative frequency column.

Monthly Wages ($)	**Number of Workers**	**Cumulative Frequency**
800–850	12	12
850–900	16	28
900–950	39	67
950–1000	56	123
1000–1050	62	185
1050–1100	75	260
1100–1150	30	290
1150–1200	10	300
Total	300	

Compute the value of Q_1 and Q_3 based on the data

Q$_1$: Find $N/4$; $N/4 = 300/4 = 75$. The cumulative frequency just greater than 75 is 123.

The first quartile class is 950–1000. Here, $l = 950$; $h = 50$; $cf = 67$; $f = 56$.

$$Q_1 = l + h \times [((N/4) - C_f)/f]$$

$$= 950 + 50 \times ((75 - 67)/56) = 950 + 50 \times 0.1429 = 950 + 7.145 = 957.145$$

The first quartile is $957.15

Q_3: Find $(3/4) \times N$; $(3/4) \times N = 225$.

The cumulative frequency just greater than 225 is 260. The third-quartile class is 1050–1100.

Here $l = 1050$; $h = 50$; $C_f = 185$; $f = 75$.

$$Q_3 = l + h \times [((3 \times (N/4)) - C_f)/f)$$

$$= 1050 + 50 \times ((225 - 185)/75) = 1050 + 50 \times 0.533 = 1050 + 26.667 = 1076.667$$

The third quartile is 1076.667

By definition, QD = $(Q_3 - Q_1)/2$; QD = (1076.67 – 957.15)/2 = 59.76

By definition, CQD = $(Q_3 - Q_1)/(Q_3 + Q_1)$

$$\text{CQD} = 119.52/2033.82 = 0.0588.$$

Hence, the quartile deviation is 59.76, and the coefficient of quartile deviation is 0.0588.

5.5 Mean Deviation

Mean deviation (MD) is the mean of the difference of each item in the distribution of its average (i.e., mean, median, mode). Consider all the deviations as a positive one (i.e., consider the magnitude of the deviations).

Discrete series

$$\textit{Mean} \text{ deviation from mean} = \sum_{i=1}^{n}\left[\frac{|X_i - \text{Mean}|}{n}\right]$$

$$\textit{Mean} \text{ deviation from median} \quad = \sum_{i=1}^{n}\left[\frac{|X_i - \text{Median}|}{n}\right]$$

$$\textit{Mean} \text{ deviation from mode} = \sum_{i=1}^{n}\left[\frac{|X_i - \text{Mode}|}{n}\right]$$

Distribution with frequency

$$\textit{Mean} \text{ deviation from mean} = \sum_{i=1}^{n}\left[\frac{|X_i - \text{Mean}| \times f_i}{n}\right]$$

$$\textit{Mean} \text{ deviation from median} = \sum_{i=1}^{n}\left[\frac{|X_i - \text{Median}| \times f_i}{n}\right]$$

$$\textit{Mean} \text{ deviation from mode} = \sum_{i=1}^{n}\left[\frac{|X_i - \text{Mode}| \times f_i}{n}\right]$$

The median is used in preference to the arithmetic mean to calculate the mean deviation because it is a minimum when the median is the point of reference. In the actual applications, arithmetic mean is used in the evaluation of MD.

NOTE: The method given can be used for continuous series by considering the mid-class interval and the frequencies.

- It is simple and easy to understand.
- It is affected by the value of each item.
- It is not used often because all the deviations are taken as positive.
- It is not suited for algebraic treatment.
- It is mostly used in dealing with small numbers of observations when there is no need for an elaborate analysis.

Example:

Consider the following continuous distribution, which relates to the sales of 100 companies.

Sales (in lakhs of $)	40–50	50–60	60–70	70–80	80–90	90–100	TOTAL
No. of Days	10	15	25	30	12	8	100

Evaluate (a) MD(Mean), (b) MD(Median), and (c) coefficient of mean deviation (CMD).

Consider the given distribution, it is continuous and of uniform length.

Sales (in lakhs of $)	**No. of Days**	**Mid-Class Interval**	$d = (X - A)/h$	**fd**	**Cumulative Frequency**
40–50	10	45	–3	–30	10
50–60	15	55	–2	–30	25
60–70	25	65	–1	–25	50
70–80	30	75	0	0	80
80–90	12	85	1	12	92
90–100	8	95	2	16	100
Total	100			–57	

Here, $h = 10$ and let $A = 75$.

$$\text{Mean} = 75 + 10 \times (-57/100) = 75 - 5.7 = 69.3 \text{ lakhs.}$$

$n/2 = 50$; cumulative frequency just greater than 50 is 80.

The median class corresponds to the class 70–80.

Here $l = 70$; $h = 10$; $cf = 50$; $f = 30$.

$$\text{Median} = 70 + 10 \times ((50 - 50)/30) = 70 + 10 \times 0 = 70 \text{ lakhs}$$

$$\text{Mean} = 69.3;\ \text{median} = 70.$$

$$\text{MD(Mean)} = 1130/100 = 11.3;\ \text{MD(Median)} = 1130/100 = 11.3$$

For the given distribution, both the deviations are one and the same.

MD(Mean) = 11.3 and MD(Median) = 11.3

To find CMD, CMD = (MD/Median) = 11.3/70 = 0.1614

X_i	f_i	$\lvert X_i - X\rvert$	$f \times \lvert X_i = X\rvert$	$\lvert X_i - \textbf{Med}\rvert$	$f \times \lvert X_i - \textbf{Med}\rvert$
45	10	24.3	243.0	25	250
55	15	14.3	214.5	15	225
65	25	4.3	107.5	5	125
75	30	5.7	171.0	5	150
85	12	15.7	188.4	15	180
95	8	25.7	205.6	25	200
	$n = 100$		1130.0		1130

Hence, MD(Mean) = 11.3 lakhs; MD(Median) = 11.3 lakhs; and CMD = 0.1614.

5.6 Standard Deviation (SD)

It is the important absolute measure of dispersion. SD is going to reveal the scatteredness of the elements in the given distribution. It is otherwise called 'root mean square deviation' from the mean. It is defined as,

DD:

$$\text{SD} = \sigma = \sqrt{\frac{\sum_{i=1}^{n}\left[X_i - \overline{X}\right]^2}{n}} \quad \text{[or]}$$

$$\text{SD} = \sigma = \sqrt{\frac{\left[\sum_{i=1}^{n} X_i^2\right]}{n} - \left[\frac{\sum_{i=1}^{n} X_i}{n}\right]^2}$$

where X_i is the different variate in the given data set [i = 1, 2, 3, …, n]. $\overline{X}$ is the mean value and n the number of variates given.

DDF:

$$n = \sum_{i=1}^{n} f_i$$

$$\text{SD} = \sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i \times \left[X_i - \overline{X}\right]^2}{n}} \quad \text{[or]}$$

$$\text{SD} = \sigma = \sqrt{\frac{\left[\sum_{i=1}^{n} f_i X_i^2\right]}{n}\left[\frac{\sum_{i=1}^{n} f_i X_i}{n}\right]^2}$$

where X_i is the different variate and f_i is their corresponding frequencies in the given data set (i = 1, 2, 3, …, n). $\overline{X}$ is the mean value and n the number of variates given.

CDF:

$$\text{SD} = \sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i \times \left[X_i - \overline{X}\right]^2}{\sum_{i=1}^{n} f_i}}$$

where X_i is the mid-class value of the ith interval, and f_i is their corresponding frequencies in the given data set (i = 1, 2, 3, …, n). $\overline{X}$ is the mean value and n the number of variates given.

$$\text{SD} = \sigma = h \times \sqrt{\frac{\sum_{i=1}^{n} f_i d_i^2}{\sum_{i=1}^{\mathrm{n}} f_i} - \left(\frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i}\right)^2}$$

where $d_i = [X_i - A]/h$ corresponds to the ith interval and f_i is their corresponding frequency in the given data set (i = 1, 2, 3, …, n). $\overline{X}$ is the mean value, and n the number of variates given.

NOTE: The square of the SD is known as the variance. It is denoted by σ^2.

Advantages

1. It is the most widely used measures of dispersion.
2. It possesses all the qualities needed for a good measure of dispersion.
3. It is not affected owing to sampling fluctuation.
4. It is the highly reliable measures of dispersion.
5. It includes all the data values and does not ignore positive and negative signs like the mean deviation.
6. It is a mathematically logical one and can be further treated mathematically.
7. It is a minimum when it is calculated from the arithmetic mean.
8. It has great practical utility in sampling, statistical inference, and fitting a normal curve.

Disadvantages

1. It is difficult to understand.
2. Its evaluation process is complicated.
3. It is not a helpful measure to compare the variability of 2 or more distributions given in different units.

Example:

Find the QD and CQD of the daily salary of 7 persons given as:

$25, $17, $19, $10, $15, $7, $12

Nature of data is discrete. Arrange the data in the ascending order 7, 10, 12, 15, 17, 19, 25

Q_1: Find $[n + 1]/4$, here it is 2; $n = 7$

Q_1 is the 2nd item. $\mathrm{Q}_1 = 10$

Q_3: Find $3 \times [n + 1]/4$, here it is 6; Q_3 is the 6th item. $Q_3 = 19$

$$\mathrm{QD} = [Q_3 - Q_1]/2 = [19 - 10]/2 = 4.5$$

$$\mathrm{CQD} = [Q_3 - Q_1]/[Q_3 + Q_1] = [19 - 10]/[19 + 10] = 0.31$$

Hence, the QD is 4.5, and the CQD is 0.31.

Example:

Find the semi-quartile range and SD of the data given:

Salary per Day ($)	10	15	20	25	30	35	40
No. of Employees	5	10	16	20	14	8	4

Nature of data is DDF.

Because the marks are in the ascending order keep it as such.

Find the total number of students and the cumulative frequency

Salary, X	**No. of Employees**	**Cumulative Frequency**	$X \times f$	$X - \overline{X}$	$[X - \overline{X}]^2$	$f \times [X - \overline{X}]^2$
10	5	5	50	–14.42	207.94	1039.68
15	10	15	150	–9.42	88.74	887.36
20	16	31	320	–4.42	19.54	312.64
25	20	51	500	0.58	0.34	6.8
30	14	65	420	5.58	31.14	435.96
35	8	73	280	10.58	111.94	895.52
40	4	77	160	15.58	242.74	970.95
Total	77		1880			4548.9

$$Q_1\text{: } [n + 1]/4 = 78/4 = 19.5$$

The cumulative frequency just greater than 19.5 is 31.

Q_1 is the value corresponding to the cumulative frequency 31, which is 20.

$$[n + 1] \times 3/4 = 19.5 \times 3 = 58.5$$

The cumulative frequency just greater than 58.5 is 65.

Q_3 is the value corresponding to the cumulative frequency 65, which is 30.

The interquartile range or QD = $[Q_3 - Q_1]/2 = 5$

$$\text{Mean} = \overline{X} = \frac{\sum_{i=1}^{n} f_i \times X_i}{\sum_{i=1}^{n} f_i} = 1880/77 = 24.42$$

$$\text{SD} = \sigma = h \times \sqrt{\frac{\sum_{i=1}^{n} f_i d_i^2}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i}\right)^2} = \left[4548.9/77\right]^{[1/2]} = 7.686 = 7.69$$

Hence the QD is \$5 and the SD is \$7.69.

Example:

Calculate the mean deviation from the following data: What light does it throw on the social conditions of the community? Also find QD and SD.

Difference (Years)	Frequency (*f*)	Difference (Years)	Frequency (*f*)
0–5	449	20–25	109
5–10	705	25–30	52
10–15	507	30–35	16
15–20	281	35–40	4

Nature of data: Continuous

Find the mid-class interval, *X*, for the difference in years.

Difference (Years)	X	f	d	$f \times d$	$f \times d^2$	$\mid X_i - \overline{X} \mid$	$f \times \mid X_i - \overline{X} \mid$	C_f
0–5	2.5	449	–3	–1347	4041	7.97	3578.53	449
5–10	7.5	705	–2	–1410	2820	2.97	2093.85	1154
10–15	12.5	507	–1	–507	507	2.03	1029.21	1661
15–20	17.5	281	0	0	0	7.03	1975.43	1942
20–25	22.5	109	1	109	109	12.03	1311.27	2051
25–30	27.5	52	2	104	208	17.03	885.56	2103
30–35	32.5	16	3	48	144	22.03	352.48	2119
35–40	37.5	4	4	16	64	27.03	108.12	2123
Total		2123		–2987	7893		11334.45	

Here, $h = 5$ and take $A = 17.5$ and $d = [X - A]/h$

$$\text{Mean} = \overline{X} = A + h \times \left\{ \sum_{i=1}^{n} [f_i \times d_i] / \sum_{i=1}^{n} [f_i] \right\}$$

$$= 17.5 + 5\ [-2987/2123] = 10.465 = 10.47 \text{ years}$$

$$\textit{Mean}\ \text{Deviation from mean} = \frac{\sum_{i=1}^{n} \mid X_i - \text{Mean} \mid \times f_i}{\sum_{i=1}^{n} [f_i]}$$

$$= 11334.45/2123 = 5.33888 = 5.34 \text{ years}$$

The inferences here are that the average difference between the age of the husband and of the wife is 5.34 years. The deviation is quite high because it is nearly half of the average.

$Q_1 = 5 + 5 \times [530.75 - 449]/705 = 5.58$, $Q_3 = 10 + 5 \times [1592.25 - 1154]/507 = 14.32$ and QD $= [Q_3 - Q_1]/2 = 4.37$ years

Standard deviation

$$\text{SD} = \sigma = h \times \sqrt{\frac{\sum_{i=1}^{n} f_i\ d_i^2}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} f_i\ d_i}{\sum_{i=1}^{n} f_i} \right)^2} = 5 \times \left\{ [7893/2123] - [-2987/2123]^2 \right\}^{[1/2]}$$

$$= 5 \times 1.319 = 6.6 \text{ years}$$

Hence, the QD and SD are 4.37 years and 6.6 years, respectively.

Example:
Evaluate the mean deviation from the median of the following data:

Class Intervals	2–4	4–6	6–8	8–10
Frequency	3	4	2	1

Construct the cumulative frequency column and [*N*/2]

Class Interval	**Frequency**	**Mid-class *X***	**Cumulative Frequency**	**$\|X - \text{Me}\|$**	**$f \times \|X - \text{Me}\|$**
2–4	3	3	3	2	6
4–6	4	5	7	0	0
6–8	2	7	9	2	4
8–10	1	9	10	4	4
Total	10				14

$$N/2 = 10/2 = 5.$$

Cumulative frequency just greater than 5 is 7; The median class is 4–6.

$$l = 4, h = 2, C_f = 3, \text{ and } f = 4$$

$$Me = 1 + h \times \left[\frac{(n/2) - C_f}{f}\right] = 4 + 2 \times \left[5 - 3\right] / 4 = 5$$

$$\textit{Mean}\text{ Deviation from median} = \frac{\sum_{i=1}^{n} |\,X_i - \text{Median}\,| \times f_i}{\sum_{i=1}^{n} [f_i]} = 14/10 = 1.4$$

The mean deviation from median is 1.4.

5.7 Relative Measures of Dispersion

Range, QD, MD, and SD are all the absolute measures of dispersion. They are all expressed in the same units based on the units of the given data set. These absolute measures are highly useful in comparing the distributions of 2 or more distributions having the same unitization. This implies that, we are in need of a measure that is helpful to study the dispersion of 1 or more distributions of different unitization. To overcome this difficulty, we seek the help of relative measures like coefficient of variation (CV), CQD, and CMD. These relative measures are independent of the units of measurement. This quality of independence helps to study the variability of 2 or more distributions.

Coefficient of Variation (CV)

As per the Karl Pearson, CV is the percentage variation in the mean. It is defined as,

$$CV = [SD/Mean] \times 100$$

It helps in comparing the populations or samples having different means and different SDs.

CQD

It is defined as, CQD = [QD/Median] × 100

CMD

CMD is defined as, CMD = [MD/Mean] × 100 or CMD = [MD/Median] × 100

Example:

In 2 industries, XYZ and ABC, engaged in the same area, the mean weekly salary (in dollars) and the SD are as follows:

Industry	Mean ($)	SD ($)	No. of Employees
XYZ	34.5	5	476
ABC	28.5	4.5	524

(a) Which industry pays more weekly salary?

(b) Which industry has greater variability in individual salary?

TS is the total salary.

Industry XYZ	Industry ABC
Mean = $\overline{X}_1 = 34.5$	Mean = $\overline{X}_2 = 28.5$
SD = $\sigma_1 = 5$	SD = $\sigma_2 = 4.5$
No. of Employees = $n_1 = 476$	$n_2 = 524$

$\overline{X}_1$ = Total salary of 476 employees of industry XYZ/476

Total salary of 476 employees of industry XYZ = 476 × $\overline{X}_1$ = 476 × 34.5 = $16,422

$$CV_1 = [\sigma_1/\overline{X}_1] \times 100 = [5/34.5] \times 100 = 14.49\%$$

$\overline{X}_2$ = Total salary of 476 employees of industry ABC/524

Total salary of 524 employees of Industry ABC = 524 × $\overline{X}_2$ = 524 × 28.5 = $14,934

$$CV_2 = [\sigma_2/\overline{X}_2] \times 100 = [4.5/28.5] \times 100 = 15.79\%$$

(a) Comparing total salary of both companies, Company XYZ pays out more salary.

(b) Because $CV_1 < CV_2$, Company ABC has greater variability in the individual salary of the employees.

Example:

Lives of 2 models of refrigerators turned in for new models in a recent survey are:

Life (Years)	No. of Refrigerators: *A*	No. of Refrigerators: *B*
0–2	5	2
2–4	16	7
4–6	13	12
6–8	7	19
8–10	5	9
10–12	4	1

What is the average life of each model of these refrigerators? Which model has more uniformity?

Consider the given data and evaluate mean and SD for *A* and *B*.

Here $A = 5$; $h = 2$; and $d = [X - A]/h$

Life (Years)	f_A	f_B	x	d	$f_A \times d$	$f_B \times d$	d^2	$f_A \times d^2$	$f_B \times d^2$
0–2	5	2	1	–2	–10	–4	4	20	8
2–4	16	7	3	–1	–16	–7	1	16	7
4–6	13	12	5	0	0	0	0	0	0
6–8	7	19	7	1	7	19	1	7	19
8–10	5	9	9	2	10	18	4	20	36
10–12	4	1	11	3	12	3	9	36	9
Total	50	50			3	29		99	79

$$\text{Mean } A = \overline{X} = A + h \times \left\{ \sum_{1}^{n} [f_i \times d_i] \Big/ \sum_{1}^{n} [f_i] \right\} = 5 + 2 \times \left[3/50 \right] = 5.12 \text{ years}$$

$$SD = \sigma = h \times \sqrt{ \frac{\sum_{i=1}^{n} f_i d_i^2}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i} \right)^2 } = 2 \times \left\{ [99/50] - [3/50]^2 \right\}^{[1/2]} = 2.81 \text{ years}$$

$$\text{CV}_\text{A}\ [\text{SD } A/\text{Mean } A] \times 100 = [2.81/5.12] \times 100 = 54.88\%$$

$$\text{Mean } B = \overline{X} = A + h \times \left\{ \sum_{i=1}^{n} [f_i \times d_i] \Big/ \sum_{i=1}^{n} [f_i] \right\}$$

$$= 5 + 2 \times [29/50] = 6.16 \text{ years}$$

$$SD = \sigma = h \times \sqrt{\frac{\sum_{i=1}^{n} f_i d_i^2}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i}\right)^2} = 2 \times \left\{[79/50] - [29/50]^2\right\}^{[1/2]}$$

$$= 2.23 \text{ years}$$

$$\text{CV}_\text{B} = [\text{SD } B/\text{Mean } B] \times 100 = [2.230/6.16] \times 100 = 36.2\%$$

Because the covariance of Model B is less than the covariance of Model A, Model B has more uniformity in life than the Model A.

Example:

A factory produces two types of electric lamps, A and B. In an experiment relating to their life, the following results were obtained:

Length of Life (Hours)	Number of Lamps: A	Number of Lamps: B
500–700	5	4
700–900	11	30
900–1100	26	12
1100–1300	10	8
1300–1500	8	6

Compare the variability of the life of the 2 varieties using CV.

Using the given data, evaluate means and standard deviations of both lamps A and B.

Here $A = 1000$ and $h = 200$

$$\text{Mean } A = A + h \times \left\{\sum_{1}^{n}[f_i \times d_i] / \sum_{1}^{n}[f_i]\right\} = 1000 + 200 \times [5/60] = 1016.67 \text{ hours}$$

$$\text{SD} = \sigma = h \times \sqrt{\frac{\sum_{i=1}^{n} f_i d_i^2}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i}\right)^2} = 200 \times \left\{[73/60] - [5/60]^2\right\}^{[1/2]}$$

$$= 219.975 \text{ hours}$$

Life (Years)	f_A	f_B	x	$d = [x - A]/h$	$f_A \times d$	$f_B \times d$	d^2	$f_A \times d^2$	$f_B \times d^2$
500–700	5	4	600	–2	–10	–8	4	20	16
700–900	11	30	800	–1	–11	–30	1	11	30
900–1100	26	12	1000	0	0	0	0	0	0
1100–1300	10	8	1200	1	10	8	1	10	8
1300–1500	8	6	1400	2	16	12	4	32	24
Total	60	60			5	–18		73	78

$$\text{CV}_{\text{A}} = [\text{SD A}/\text{Mean A}] \times 100 = [219.975/1016.67] \times 10 = 21.6368\%$$

$$\text{Mean B} = A + h \times \left\{ \sum_{1}^{n} [f_i * d_i] / \sum_{1}^{n} [f_i] \right\} = 1000 + 200 \times [-18/60] = 940 \text{ hours}$$

$$\text{SD} = \sigma = h \times \sqrt{ \frac{\sum_{i=1}^{n} f_i d_i^2}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i} \right)^2 } = 200 \times \left\{ [78/60] - [-18/60]^2 \right\}^{[1/2]} = 220 \text{ hours}$$

$$\text{CV}_{\text{B}} = [\text{SD B}/\text{Mean B}] \times 100 = [220/940] \times 100 = 23.4\%$$

Because the covariance of lamp A is less than the covariance of lamp B, lamp A is better than lamp B.

Example:
The CV of 2 series are 60 and 80, respectively. Their SDs are 9 and 16, respectively. What are their arithmetic means?

Given

Mean = $\overline{X}_1$ = ?	Mean = $\overline{X}_2$ = ?
SD = $\sigma_1 = 9$	SD = $\sigma_2 = 16$
$\text{CV}_1 = [\sigma_1/\overline{X}_1] \times 100 = 60$	$\text{CV}_2 = [\sigma_2/\overline{X}_2] \times 100 = 80$
$[9/\overline{X}_1] \times 100 = 60$	$[16/\overline{X}_2] \times 100 = 80$
$\overline{X}_1 = [9/60] \times 100 = 15$	$\overline{X}_2 = [16/80] \times 100 = 20$

Hence, the means of the two series is 15 and 20.

Exercise 5

1. Calculate the average deviation from the mean, SD, and CV for the following series.

Sales	**Number of Days**
40–50	10
50–60	15
60–70	25
70–80	30
80–90	12
90–100	8

2. An analysis of the monthly wages gives the following:

	Firm A	Firm B
Number of Workers	500	600
Variances	81	100
Average Wages	$ 186	$ 175

   a. Which firm pays out a larger wage bill?
   b. In which firm does greater variability occur?
   c. What is the average if firm A and firm B are combined?

3. There are a number of possible measures of sales performance, including how consistent a sales person is in meeting established goals. The data that follow represent the percentage of goal met by each of the 3 sales persons the last 5 years:

Salesperson, X	88	68	89	92	103
Salesperson, Y	76	88	90	86	79
Salesperson, Z	104	88	118	88	123

   Which sales person is the most consistent?

4. Particulars regarding weekly wages paid to workers in firms A and B are given:

	Firm A	Firm B
Average Weekly Wages ($)	525	475
Variance of Wages ($)	100	121

   In which firm is the variation in individual wages greater?

5. Prices of a particular commodity in 5 years in 2 cities are given:

City A (Price, $)	20	22	19	23	26
City B (Price, $)	10	20	18	12	15

   Find which city had more stable prices.

6. For the data give, compute the value of QD.

Monthly Wages ($)	No. of Workers
Below 850	12
850–900	16
900–950	39
950–1000	56
1000–1050	62
1050–1100	75
1100–1150	30
1150 and Above	10

   (Hint: Convert the first and the last intervals as 800–850 and 1150–1200.)

7. Consider the following grouped data that relate to the profits of 100 companies.

Sales (in lakhs of $)	8–10	10–12	12–14	14–16	16–18	18–20
No. of Days	8	12	20	30	20	10

   Find the SD.

8. The following data show the daily sales at a petrol station. Calculate the mean, SD, and CV.

Number of Litres Sold	**Number of Days**	**Number of Litres Sold**	**Number of Days**
700–1000	12	1900–2200	18
1000–1300	18	2200–2500	5
1300–1600	20	2500–2800	2
1600–1900	15		

6

Skewness, Moments, and Kurtosis

6.1 Introduction

Skewness refers to the lack of symmetry of a distribution. A symmetrical distribution will look like a bell-shaped curve. If a distribution is not symmetrical, it is called 'skewed'. The value of mean, median, and mode will be exactly the same for a symmetrical distribution. Any difference in the values of the 3 measures clearly spells out that the corresponding distribution is skewed (not symmetric). In probability theory and statistics, 'kurtosis' (from the Greek word *kyrtos* or *kurtos*, meaning 'bulging') is a measure of the 'peakedness' of the probability distribution of a real-valued random variable.

6.2 Dispersion and Skewness

The measure dispersion shows the scatter of the items. It means that how much each item differs from the mean, whereas the skewness measures the degree of symmetry of the distribution.

Measure of skewness
Karl Pearson's measure of skewness for a skewed distribution

$$S_k = [\text{Mean} - \text{Mode}]/\text{SD}$$

For a moderately skewed distribution

$$S_k = 3 \times [\text{Mean} - \text{Median}]/\text{SD}$$

Bowley's measure of skewness

$$S_k = \{Q_3 - 2 \times Q_2 + Q_1\}/\{Q_3 - Q_1\}$$

Both the measures can be used to evaluate the skewness. If the frequency distribution has open-ended classes, Bowley's measure is the best one.

It can be classified into positively skewed and negatively skewed.

Nature	Characteristics
Positively Skewed	Mean > Median > Mode Tail of the curve extends longer to the right
Negatively Skewed	Mode > Median > Mean Tail of the curve extends to the left

It is not only necessary to find only the skewness, but also its direction. This is why 2 different distributions can have the same mean and standard deviation, but they may be skewed in the opposite direction.

Example:

Calculate the coefficient of skewness based on the mean and median of the following distribution:

Salary ($)	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80

No. of Employees	6	12	22	48	56	32	18	6

Based on the given data, evaluate the mean, median, and standard deviation.

Salary	**Employees**	C_f	**Mid** X	d	$f \times d$	d^2	$f \times d^2$
0–10	6	6	5	–4	–24	16	96
10–20	12	18	15	–3	–36	9	108
20–30	22	40	25	–2	–44	4	88
30–40	48	88	35	–1	–48	1	48
40–50	56	144	45	0	0	0	0
50–60	32	176	55	1	32	1	32
60–70	18	194	65	2	36	4	72
70–80	6	200	75	3	18	9	54
Total	200				–66		498

Let $A = 45$; $h = 10$; $n = 200$; and $d = [X - A]/h$

$$\text{Mean} = \overline{X} = A + h \times \left\{ \sum_{i=1}^{n} [f_i \times d_i] / \sum_{i=1}^{n} [f_i] \right\}$$

$$= 45 + 10 \times \left[-66/200 \right] = 41.7$$

$n/2 = 100$, cumulative frequency just greater than 100 is 144, the median class is 40–50.

$C_f = 88, f = 56$; and $l = 40$

$$Mean = 1 + h \times \left[\frac{(n/2) - C_f}{f} \right] = 40 + 10 \times \left[\left\{ 100 - 88 \right\} / 56 \right] = 42.14$$

$$\mathrm{SD} = \sigma = h \times \sqrt{\frac{\sum_{i=1}^{n} f_i d_i^{\,2}}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i}\right)^2} = 10 \times [2.49 - 0.1089]^{[1/2]} = 15.43$$

$S_k = 3 \times [\text{Mean} - \text{Median}]/\text{SD} = 3 \times [41.7 - 42.14]/15.43$

$= -0.0855$

Example:

Calculate Pearson's measure of skewness on the basis of mean, mode, and standard deviation from the following data.

Class Interval	14–15	15–16	16–17	17–18	18–19	19–20	20–21	21–22
Frequency	35	40	48	100	125	87	43	22

Based on the given data, evaluate mean, mode and standard deviation.

$$A = 18.5,\ d = [X - A]/h,\ h = 1$$

$$\text{Mean} = \overline{X} = A + h \times \left\{ \sum_{i=1}^{n} [f_i \times d_i] \ / \ \sum_{i=1}^{n} [f_i] \right\}$$

$$= 18.5 + 1 \times [-217/500] = 18.5 - 0.434 = 18.066$$

Marks	**No. of Students**	**Mid *X***	***d***	***f* × *d***	***d*²**	***f* × *d*²**
14–15	35	14.5	–4	–140	16	560
15–16	40	15.5	–3	–120	9	360
16–17	48	16.5	–2	–96	4	192
17–18	100	17.5	–1	–100	1	100
18–19	125	18.5	0	0	0	0
19–20	87	19.5	1	87	1	87
20–21	43	20.5	2	86	4	172
21–22	22	21.5	3	66	9	198
Total	500			–217		1669

The maximum frequency is 125. It implies that the modal class is 18–19

$$\text{Mode} = 1 + h * \left[\frac{(f_0 - f_1)}{(2f_0 - f_1 - f_2)} \right]$$

$1 = 18, f_0 = 125, f_1 = 100$, and $f_2 = 87$

Mode = 18 + 1 × [(125 − 100)/(2 × 125 – 100 − 87)] = 18 + 0.3968 = 18.3968

$$\text{SD} = \sigma = h \times \sqrt{\frac{\sum_{i=1}^{n} f_i d_i^2}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i}\right)^2} = 1\left\{[1669/500] - [-217/500]^2\right\}^{[1/2]}$$

$$= 1.7747 = 1.78$$

Skewness = 3 × [Mean – Mode]/SD

$$= 3 \times [18.066 - 18.3968]/1.78 = -0.5575 = -0.56$$

Hence, mean = 18.066, mode = 18.3968, SD = 1.78, and skewness = −0.56

Example:

Which group is more symmetrically skewed?

Group	Mean	Median	SD
I	22	24	10
II	22	25	12
Group I			Group II
Mean = 22			Mean = 22
Median = 24			Median = 25
SD = 10			SD = 12

Skewness = 3 × [Mean – Median]/SD

Skewness for Group I: $S_k = 3 \times [22 - 24]/10 = -0.6$

Skewness for Group II: $S_k = 3 \times [22 - 25]/12 = -0.75$

Because the absolute value of the skewness of Group I is less than Group II, Group I is more symmetrical than Group II.

6.3 Moments

Moments are used to refer to the peculiarities of a frequency distribution. The utility of moments lies in the sense that they indicate different aspects of a given distribution. They help to measure the central tendency of a series, dispersion or variability, skewness, and the peakedness of the curve.

The moments about the actual arithmetic mean are denoted by the symbol, μ. The first 4 moments about the mean are as follows:

$$\mu_1 = \frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]$$

$$\mu_2 = \frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^2$$

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^3$$

$$\mu_4 = \frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^4$$

In general the *r*th moment can be defined as

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^r$$

NOTE: The values of *x* are discrete in nature.

If the values of *x* are referring to frequency distributions, then the formulas for the 4 moments are given as follows:

$$\mu_1 = \frac{1}{n}\sum_{i=1}^{n}f_i[x_i - \bar{x}]$$

$$\mu_2 = \frac{1}{n}\sum_{i=1}^{n}f_i[x_i - \bar{x}]^2$$

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{n}f_i[x_i - \bar{x}]^3$$

$$\mu_4 = \frac{1}{n}\sum_{i=1}^{n}f_i[x_i - \bar{x}]^4$$

In general the *r*th moment can be defined as

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n}f_i[x_i - \bar{x}]^r$$

6.4 Kurtosis

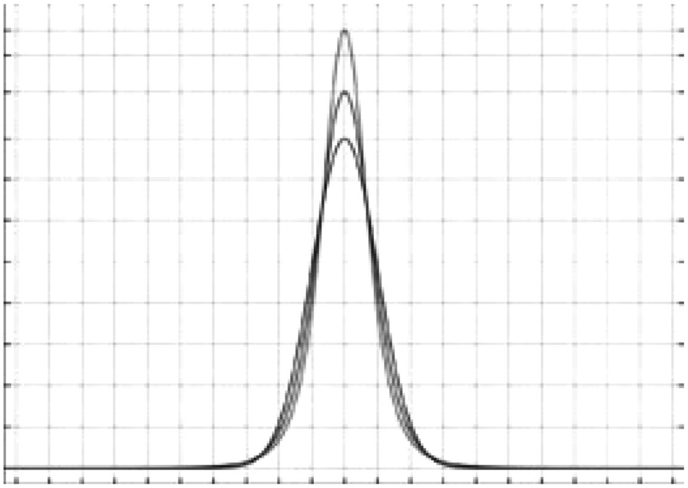
Kurtosis refers to the peakedness of the frequency curve. Pearson (1905) introduced kurtosis as a measure of how flat the top of a symmetric distribution is when compared to a normal distribution of the same variance. According to Clark, the term 'kurtosis' means the property of the distribution that expresses the peakedness. It is denoted by the notation β_2.

$$\beta_2 = [\mu_4]/[\mu_2]^2$$

NOTE: The value of skewness can also be evaluated using moments. Skewness = $[\mu_3]/[\mu_2]^{1.5}$

A high kurtosis distribution has a sharper 'peak' and fatter 'tails', whereas a 'low kurtosis distribution has a more rounded peak with wider shoulders'.

Distributions with 0 kurtosis are called 'mesokurtic' ($\beta_2 = 3$). The most prominent example of a mesokurtic distribution is the normal distribution family, regardless of the values of its parameters.



A distribution with positive kurtosis is called 'leptokurtic' ($\beta_2 > 3$). In terms of shape, a leptokurtic distribution has a more acute 'peak' around the mean (that is, a higher probability than a normally distributed variable of values near the mean) and 'fat' tails.

Examples of leptokurtic distributions include the Laplace distribution and the logistic distribution. Such distributions are sometimes termed 'super Gaussian'.

A distribution with negative kurtosis is called 'platykurtic' ($\beta_2 < 3$). In terms of shape, a platykurtic distribution has a smaller 'peak' around the mean and 'thin tails'. Examples of platykurtic distributions include the continuous or discrete uniform distributions and the raised cosine distribution. The most platykurtic distribution of all is the Bernoulli distribution with $p = ½$.

Example:

Evaluate the first 4 moments of the following discrete distribution:

45	55	65	75	85

Hence, evaluate the value of kurtosis.

Step 1: First evaluate the mean value.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 325/5 = 65$$

Step 2: The required four moments are

$$\mu_1 = \frac{1}{5}\sum_{i=1}^{5}[x_i - 65] = 0$$

$$\mu_2 = \frac{1}{5}\sum_{i=1}^{5}[x_i - 65]^2 = 200$$

$$\mu_3 = \frac{1}{5}\sum_{i=1}^{5}[x_i - 65]^3 = 0$$

$$\mu_4 = \frac{1}{5}\sum_{i=1}^{5}[x_i - 65]^4 = 68000$$

$$\text{Kurtosis} = \beta_2 = [\mu_4]/[\mu_2]^2$$

$$\beta_2 = 68000/[200]^2 = 1.7$$

NOTE: The value of β_2 is 1.7, which is less than 3, and implies that the given distribution is platy-kurtic.

Example:

Find the moments of the following data

X	1	2	3	4	5	6	7	8	9
Frequency	1	6	13	25	30	22	9	5	2

Also evaluate the value of kurtosis.

Mean = 4.912; n = 113

$$\mu_1 = \frac{1}{113}\sum_{i=1}^{9} f_i[x_i - 4.91] = 0$$

$$\mu_2 = \frac{1}{113}\sum_{i=1}^{9} f_i[x_i - 4.91]^2 = 2.49$$

$$\mu_3 = \frac{1}{113}\sum_{i=1}^{9} f_i[x_i - 4.91]^3 = 0.68$$

$$\mu_4 = \frac{1}{113}\sum_{i=1}^{9} f_i[x_i - 4.91]^4 = 18.34$$

Kurtosis = $\beta_2 = [\mu_4]/[\mu_2]^2 = \beta_2 = 18.34/[2.49]^2 = 2.96$

Example:

Find all the four moments for the following data

Class Interval	**0–10**	**10–20**	**20–30**	**30–40**
Frequency	1	3	4	2

Hence, evaluate the value of kurtosis.

Here the v3alue of x refers to the mid-value of the class intervals.

Mean = 22; m = 10

$$\mu_1 = \frac{1}{10}\sum_{i=1}^{4} f_i[x_i - 22] = 0$$

$$\mu_2 = \frac{1}{10}\sum_{i=1}^{4} f_i[x_i - 22]^2 = 81$$

$$\mu_3 = \frac{1}{10}\sum_{i=1}^{4} f_i[x_i - 22]^3 = -144$$

$$\mu_4 = \frac{1}{10}\sum_{i=1}^{4} f_i[x_i - 22]^4 = 14817$$

Kurtosis = $\beta_2 = [\mu_4]/[\mu_2]^2$

$\beta_2 = 14817/[81]^2 = 2.258 = 2.26$

Exercise 6

1. In a frequency distribution, the coefficient of skewness (given by Bowley) based on the quartiles is 0.6. If the sum of the upper and lower quartiles is 100 and median is 38, find the value of the upper and lower quartiles.

   (NOTE: given median = $Q_2 = 38$; $S_k = 0.6$; $Q_1 + Q_3 = 100$; use it in the formula related to S_k; implies that $Q_3 - Q_1 = 1.33$; solve for Q_1 and Q_3.)

2. In a hotel with 70 rooms, data is collected for 104 days on the number of rooms occupied on a day. This data is classified and is given:

No. of Rooms Occupied	No. of Days
0–10	10
10–20	12
20–30	18
30–40	25
40–50	16
50–60	15
60–70	8

   Compute Karl Pearson's coefficient of skewness and interpret its value. Also evaluate the value of kurtosis.

3. Calculate kurtosis for the following distribution:

Size	25	35	45	55	65	75	85
Frequency	3	61	132	153	140	51	2

4. Find the value of kurtosis for the following data:

Expenditure Per month (Rs. in 000)	3–6	6–9	9–12	12–15	15–18	18–21	21–24
Number of Families	28	292	389	212	59	18	2

5. Assume that a firm has selected a random sample of 100 from its production line and collected the data as follows:

Class	130–134	135–139	140–144	145–149	150–154	155–159	160–164
Freq.	3	12	21	28	19	12	5

   Compute the mean, SD, skewness, and kurtosis.

7

Correlation and Regression Analysis

7.1 Introduction

We shall now study 2 (bivariate) or more variables (multivariate) simultaneously and attempt to find the relationship between the variables in quantitative or qualitative form. We have many such related variables, like crops per acre and fertilizer, height and weight, income and expenditure, etc.

The credit of this methodology of studying the strength of relationships among the variables goes to Sir Francis Galton and Karl Pearson.

7.2 Correlation

Correlation is a statistical measure used to evaluate the strength and degree of the relationship among the 2 or more variables under study. Here the term, 'relationship' is used to measure the tendency of the variables to move together. The movement of the variables may be in the same direction or in the opposite direction. The correlation is said to be positive if the variables are moving in the same direction and negative, if they are moving in the opposite direction. If there is no change in direction, it implies that the variables are not related.

Correlation is classified into

- simple correlation,
- rank correlation, and
- group correlation.

7.2.1 Simple Correlation or Correlation

This measure can be evaluated in a discrete series of quantitative in nature. It is denoted by the notation r. The value of r lies in the closed interval ($-1 \leq r \leq 1$). If the value of r is toward 1, then variables are said to be positively correlated or directly related (if X increases, Y also increases, and if X decreases, Y also decreases). If it is toward -1, then it is said to be negatively correlated or inversely related (if X increases, Y will decrease, and if X decreases, Y increases), and if it is 0, then the variables are said to be uncorrelated (the change in X does not affect the variable Y and vice versa).

7.2.2 Rank Correlation

Rank correlation can be evaluated in a discrete series of qualitative in nature. It is denoted by R. The value of R lies in the closed interval ($-1 \leq R \leq 1$).

7.2.3 Group Correlation

Group correlation measure can be evaluated in a continuous series of grouped data. It is denoted by r. The values of r lie in the closed interval ($-1 \leq r \leq 1$).

NOTE: The larger the value of r, the stronger the linear relationship between Y and X. If $r = -1$ or $r = +1$, the regression line will include all data points and the line will be a perfect fit.

7.2.4 Assumptions for Karl Pearson's Coefficient of Correlation

- The relationship between the two series (X and Y) is linear (the amount of variation in X bears a constant ratio to the corresponding amount of variation in Y).
- Either one of the series is dependent on the other, or both are depending on the third series.
- Correlation analysis is applied to most scientific data where inferences are to be made. In agriculture, amount of fertilizer and crop yields are correlated; in economics, prices and demand or money and prices are correlated. In medicine, use of cigarettes and incidence of lung cancer or use of new drug and the percentage of cases cured are correlated. In sociology, unemployment and crime or welfare expenditure and labour efficiency are correlated. And in demography, wealth and fertility and so on are correlated.
- The correlation coefficient, r, like other statistics of the sample, is tested to see how the sample results may be generalized to the parent population.

7.2.5 Limitations

- Interpretation of this analysis needs expertise regarding the statistical concepts and the background data.
- Correlation in statistics is studied by scatter diagrams, regression lines, or coefficient of correlation.

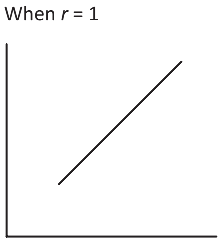
7.2.6 Properties

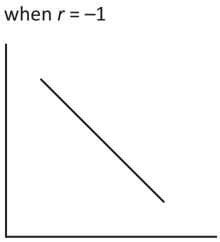
- It is independent of any change of origin of reference and the units of measurement.
- Its value lies in the interval [–1, 1].
- It is a constant value, which helps to measure the relationship between 2 variables.

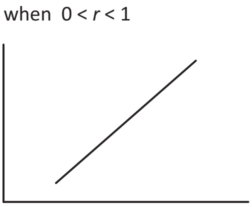
7.2.7 Scatter Diagram

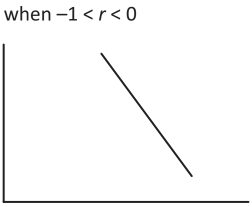
The scatter diagram is a valuable graphic device to show the existence of correlation between the two variables. Represent the variable X on the x-axis and Y on the y-axis.

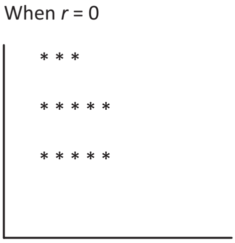
Mark the coordinate's points (x, y), and then the existence of correlation can be studied based on the structure of the clustering of the coordinate's points. The direction of scatter reveals the refuse and the strength of the scatter correlation between the variables.











The scatter diagrams for r and $0 < r < 1$ indicates that the path is linear, and the variables are moving in the same direction. This indicates the correlation is positive because the relationship between the variables is direct.

The scatter diagrams for $Y = -1$ and $-1 < Y < 0$ indicate that the variables are moving in opposite directions and the path is linear.

The scatter diagram for $Y = 0$ indicates that the variables are not related, and the path is a curve.

7.3 Karl Pearson Coefficient of Correlation

Consider the pairs of values (X_1, Y_1), (X_2, Y_2), … , (X_n, Y_n) of the variables X and Y. Then the covariance of these 2 variables X and Y can be defined as

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}[X_i - \bar{X}][Y_i - \bar{Y}]}{n}$$

The standard deviations of X and Y can be given by

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}[X_i - \bar{X}]^2}{n}}$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{n}[Y_i - \bar{Y}]^2}{n}}$$

The correlation coefficient r can be defined as

$$r = \frac{Cov(X, Y)}{\sigma_x \ \sigma_y}$$

Equivalent alternative formulas for r:

1. $r = \dfrac{\sum_{i=1}^{n}[X_i - \bar{X}][Y_i - \bar{Y}]}{\sqrt{\sum_{i=1}^{n}[X_i - \bar{X}]^2 \sum_{i=1}^{n}[Y_i - \bar{Y}]^2}}$

2. $r = \dfrac{\left[\dfrac{\sum_{i=1}^{n}[X_i \, Y_i]}{n}\right] - \left[\dfrac{\sum_{i=1}^{n} X_i}{n}\right]\left[\dfrac{\sum_{i=1}^{n} Y_i}{n}\right]}{\sqrt{\left\{\sum_{i=1}^{n}\left(\dfrac{X_i^2}{n}\right) - \bar{X}^2\right\}}\sqrt{\left\{\sum_{i=1}^{n}\left(\dfrac{Y_i^2}{n}\right) - \bar{Y}^2\right\}}}$

Value of r using assumed mean

To derive the result, we make use of the concept that the correlation coefficient is independent of choice of origin. Take $X_i = [X - a]$ and $Y_i = [Y - b]$, where a is any value of X and b is any value of Y. Then

$$r = \frac{\sum_{i=1}^{n}[X_i - a][Y_i - b]}{\sqrt{\left[\sum_{i=1}^{n}[X_i - a]^2 \times \sum_{i=1}^{n}[Y_i - b]^2\right]}}$$

Example:

In trying to evaluate the effectiveness of its advertising campaign, a firm compiled the following information:

Year	Advertisement Expenditure ($000)	Sales ($ in Lakhs)
1998	12	5.0
1999	15	5.6
2000	15	5.8
2001	23	7.0
2002	24	7.2
2003	38	8.8
2004	42	9.2
2005	48	9.5

Find the value of *r*.

Let the variables *X* and *Y* refer advertising expenditure (1 unit = $1000) and sales (1 unit = $100,000), respectively.

Let *a* = 23 and *b* = 7.0

X	Y	$X-a$	$Y-b$	$(X-a)(Y-b)$	$(X-a)^2$	$(Y-b)^2$
12	5	−11	−2	22	121	4
15	5.6	−8	−1.4	11.2	64	1.96
16	5.8	−7	−1.2	8.4	49	1.44
23	7	0	0	0	0	0
24	7.2	1	0.2	0.2	1	0.04
38	8.8	15	1.8	27	225	3.24
42	9.2	19	2.2	41.8	361	4.84
48	9.5	25	2.5	62.5	625	6.25
				173.1	1446	21.77

$$r = \frac{\sum_{i=1}^{n}[X_i - a][Y_i - b]}{\sqrt{\sum_{i=1}^{n}[X_i - a]^2 \sum_{i=1}^{n}[Y_i - b]^2}}$$

$$= \frac{173.1}{\sqrt{1446 \times 21.77}}$$

$$= 0.9756$$

The advertisement expenditure and the sales level are positively related with correlation 0.9756.

Example:

The personnel manager of an electronic manufacturing company devises a manual dexterity tests for job applicants to predict their production rating in the assembly department. To do this he selects a random sample of 10 applicants. They are given the test and later assigned a production rating. The results are as follows:

Worker	Test Scores	Production Rating
A	53	45
B	36	43
C	88	89
D	84	79
E	86	84
F	64	66
G	45	49
H	48	48
I	39	43
J	69	76

Determine the correlation coefficient.

Let the variables X and Y refer to the test score and the production rating, respectively.

Worker	X	Y	XY	X^2	Y^2
A	53	45	2385	2809	2025
B	36	43	1548	1296	1849
C	88	89	7832	7744	7921
D	84	79	6636	7056	6241
E	86	84	7224	7396	7056
F	64	66	4224	4096	4356
G	45	49	2205	2025	2401
H	48	48	2304	2304	2304
I	39	43	1677	1521	1849
J	69	76	5244	4761	5776
Total	612	622	41279	41008	41778

Here, $n = 10$.

$$r = \frac{\left[\frac{\sum_{i=1}^{n}[X_i\,Y_i]}{n}\right] - \left[\frac{\sum_{i=1}^{n}X_i}{n}\right]\left[\frac{\sum_{i=1}^{n}Y_i}{n}\right]}{\sqrt{\left\{\sum_{i=1}^{n}\left(\frac{X_i^2}{n}\right) - \bar{X}^2\right\}}\sqrt{\left\{\sum_{i=1}^{n}\left(\frac{Y_i^2}{n}\right) - \bar{Y}^2\right\}}}$$

$$= \frac{\left[\frac{41279}{10}\right] - \left[\frac{612}{10}\right]\left[\frac{622}{10}\right]}{\sqrt{\frac{41008}{10} - \left(\frac{612}{10}\right)^2}\sqrt{\frac{41778}{10} - \left(\frac{622}{10}\right)^2}}$$

$$= 0.97$$

The correlation value is 0.97, and it implies that the test score and the production rate of the employees have a close positive relation.

Example:

Calculate Karl Pearson's coefficient of correlation. From the following data using 20 as the working mean for the price and 70 as the working mean for demand.

Price ($)	14	16	17	18	19	20	21	22	23
Demand	84	78	70	75	66	67	62	58	60

Let the variables *X* and *Y* refer the price and demand, respectively.

The assumed means are given as $a = 20$ and $b = 70$

Price, *X*	**Demand, *Y***	***X* – *a***	***Y* – *b***	**(*X* – *a*)(*Y* – *b*)**	**(*X* – *a*)²**	**(*Y* – *b*)²**
14	84	–6	14	–84	36	196
16	78	–4	8	–32	16	64
17	70	–3	0	0	9	0
18	75	–2	5	–10	4	25
19	66	–1	–4	4	1	16
20	67	0	–3	0	0	9
21	62	1	–8	–8	1	64
22	58	2	–12	–24	4	144
23	60	3	–10	–30	9	100
Total				–184	80	618

Here, $n = 9$

$$Y = \frac{\sum_{i=1}^{n}[X_i - a][Y_i - b]}{\sqrt{\sum_{i=1}^{n}[X_i - a]^2 \sum_{i=1}^{n}[Y_i - b]^2}}$$

$$= \frac{-184}{\sqrt{80 \times 618}}$$

$$= -0.827520$$

$$= -0.83$$

The correlation value, –0.83, implies that the demand and the price are negatively related.

Example:

A computer, while calculating the value Y between 2 variables X (i.e., advertising expenditure) and Y (sales level) from 25 sets of values gives $n = 25$; $\Sigma X = 125$; $\Sigma Y = 100$; $\Sigma X^2 = 650$; $\Sigma Y^2 = 460$; and $\Sigma XY = 508$. It was found that 2 sets of values were wrongly entered.

Wrong Value		Correct Value	
X	*Y*	*X*	*Y*
6	14	8	12
8	6	6	8

Evaluate the correct value of r.

Given,

$n = 25$; $\Sigma X = 125$; $\Sigma Y = 100$; $\Sigma X^2 = 650$; $\Sigma Y^2 = 460$; and $\Sigma XY = 508$. First, we have to find the corrected sums that is, subtract the incorrect values and add the correct values from the total.

Corrected Values:

$$\Sigma X = 125 - (\text{sum of incorrect values}) + (\text{sum of correct values})$$

$$\Sigma X = 125 - (6 + 8) + (8 + 6) = 125 - 14 + 14 = 125.$$

Similarly proceeding,

$$\Sigma Y = 100 - (14 + 6) + (12 + 8) = 100 - 20 + 20 = 100;$$

$$\Sigma X^2 = 650 - (6^2 + 8^2) + (8^2 + 6^2) = 650;$$

$$\Sigma Y^2 = 460 - (14^2 + 6^2) + (12^2 + 8^2) = 460 - 232 + 208 = 436;$$

$$\Sigma XY = 508 - (6 \times 14 + 8 \times 6) + (8 \times 12 + 6 \times 8) = 508 - (132 + 144) = 520$$

$$r = \frac{\left[\dfrac{\sum_{i=1}^{n}[X_i\ Y_i]}{n}\right] - \left[\dfrac{\sum_{i=1}^{n} X_i}{n}\right]\left[\dfrac{\sum_{i=1}^{n} Y_i}{n}\right]}{\sqrt{\left\{\sum_{i=1}^{n}\left(\dfrac{X_i^2}{n}\right) - \bar{X}^2\right\}}\sqrt{\left\{\sum_{i=1}^{n}\left(\dfrac{Y_i^2}{n}\right) - \bar{Y}^2\right\}}}$$

$$= \frac{(520/25) - (125/25) \times (100/25)}{\sqrt{[(650/25) - ((125/25)^2)] \times [(436/25) - (100/25)^2]}}$$

$$= 0.8/1.44$$

$$= 0.56$$

Hence, the corrected value of the correlation coefficient is 0.56.

7.4 Coefficient of Correlation of a Grouped Data

In grouped data, the information is given in a correlation table. In each compartment of the table, the deviations from the average of x and the average of y with respect to the corresponding compartment are multiplied and put atop the actual figure. This outcome is further multiplied by that figure and added to give Σxy.

$$r = \frac{\left[\dfrac{\sum_{i=1}^{n}[f_i\ dx\ dy]}{n}\right] - \left[\dfrac{\sum_{i=1}^{n} f_i\ dx}{n}\right]\left[\dfrac{\sum_{i=1}^{n} f_i\ dy}{n}\right]}{\sqrt{\sum_{i=1}^{n}\left(\dfrac{f_i\ dx^2}{n}\right) - \left(\dfrac{\sum_{i=1}^{n} f_i\ dx}{n}\right)^2}\ \sqrt{\sum_{i=1}^{n}\left(\dfrac{f_i\ dy^2}{n}\right) - \left[\dfrac{\sum_{i=1}^{n} f_i\ dy}{n}\right]^2}}$$

Example:

The following table gives the distribution of total population and those who are totally are partially blind among them. Find if there is any relation between age and blindness.

Age	No. of Persons in 000	Blind
0–10	100	45
10–20	60	40
20–30	40	40
30–40	36	40
40–50	24	36
50–60	11	22
60–70	6	18
70–80	3	15

Create a modified table that comprised the data, percentage of blindness over the population.

$$Y = \text{ratio of blind} = \frac{\text{Number of Blind}}{\text{Number of Persons}}$$

Age	**Mid-Class,** x	$d_x = x - A/h$	d_x^2	y **(Ratio-Blind)**	$d_y = y - 1.5$	d_y^2	$d_x d_y$
0–10	5	−4	16	0.45	−1.05	1.1	4.2
10–20	15	−3	9	0.67	−0.83	0.69	2.49
20–30	25	−2	4	1	−0.5	0.25	1
30–40	35	−1	1	1.11	−0.39	0.15	0.39
40–50	45	0	0	1.5	0	0	0
50–60	55	1	1	2	0.5	0.25	0.5
60–70	65	2	4	3	1.5	2.25	3
70–80	75	3	9	5	3.5	12.25	10.5
		−4	44		2.73	16.94	22.08

Let $A = 45$; $h = 10$; $n = 8$.

$$r = \frac{n\sum dxdy - \left[\sum dx\right]\left[\sum dy\right]}{\sqrt{\left(n\sum dx^2 - \left[\sum dx\right]^2\right)}\sqrt{\left(n\sum dy^2 - \left[\sum dy\right]^2\right)}} = \frac{8 \times 22.08 - [-4] \times 2.73}{\sqrt{8 \times 44 - [-4]^2}\sqrt{8 \times 16.94 - [2.73]^2}}$$

$$r = \frac{187.56}{\sqrt{43030.54}} = 0.90$$

There is a close positive coordination between age and blindness.

Example:

Find the coefficient of correlation between the ages of husbands and the ages of wives given in the form of a 2-way frequency table:

Age of husbands (in years)

Ages of Husbands (Years) **Ages of Wives (Years)**	**20–25**	**25–30**	**30–35**	**35–40**	**Total**
15–20	20	10	3	2	35
20–25	4	28	6	4	42
25–30	—	5	11	—	16
30–35	—	—	2	—	2
35–40	—	—	—	—	0
Total	24	43	22	6	95

Let X refer mid-class interval of age of husbands in years

Y refers mid-class interval of age of wives in years

$$h = 5;\ dx = X - A/h;\ dy = X - B/h.$$

$$\Sigma fdxdy = 128;\ \Sigma fdx = -85;\ \Sigma fdy = -110$$

$$\Sigma fdx^2 = 145;\ \Sigma fdy^2 = 184;\ n = \Sigma f = 95.$$

$$r = \frac{n\sum fdxdy - \left[\sum fdx\right]\left[\sum fdy\right]}{\sqrt{\left(n\sum fdx^2 - \left[\sum fdx^2\right]\right)}\sqrt{\left(n\sum fdy^2 - \left[\sum fdy^2\right]\right)}} = \frac{95\times128-(-85)(-110)}{\sqrt{95\times145-(-85)^2}\sqrt{95\times184-(-110)^2}}$$

$$r = \frac{2810}{5936.24} = 0.47$$

Class Interval (Men)			20–25	25–30	30–35	35–40				
		X	22.5	27.5	32.5	37.5				
Class Interval (Women)	***Y***	*dy* / *dx*	−2	−1	0	1	**Total, *f***	***fdy***	***fd²y***	***fdxdy***
15–20	17.5	−2	20(4)	10(2)	3(0)	2(−2)	35	−70	140	96
20–25	22.5	−1	4(2)	28(1)	6(0)	4(−1)	42	−42	42	32
25–30	27.5	0	—	5(0)	11(0)	—	16	0	0	0
30–35	32.5	1	—	—	2(0)	—	2	2	2	0
35–40	37.5	2	—	—	—	—	—	0	0	
		Total, *f*	24	43	22	6	95	−110	184	128
		fdx	−48	−43	0	6	−85			
		fd²x	96	43	0	6	145			
		fdxdy	88	48		−8	128			

Example:

Show that r lies between +1 and −1.

Let $X_i = X_i - \bar{X}$ and let $Y_i = Y_i - \bar{Y}$

$$\text{Consider } \Sigma X_i^2 \times \Sigma Y_i^2 - (\Sigma X_iY_i)^2 = (X_1^2 + X_2^2 + \ldots + X_n^2)(Y_1^2 + Y_2^2 + \ldots + Y_n^2) - (X_1Y_1 + X_2Y_2 + \ldots + X_nY_n)^2.$$

$$= (X_1^2Y_1^2 + X_2^2Y_2^2 + \ldots + X_1^2Y_n^2 + X_2^2Y_1^2 + X_2^2Y_2^2 + \ldots + X_n^2Y_1^2 + X_n^2Y_2^2 + \ldots + X_n^2Y_n^2) - (X_1^2Y_1^2 + X_2^2Y_2^2 + \ldots + X_n^2Y_n^2 + 2X_1Y_1X_2Y_2 + \ldots)$$

$$= (X_1^2Y_2^2 + X_2^2Y_1^2 - 2X_1Y_2X_2Y_1) + (X_1^2Y_3^2 + X_3^2Y_1^2 - 2X_1Y_3X_3Y_1) + \ldots \tag{7.1}$$

$$= (X_1Y_1 - X_2Y_1)^2 + (X_1Y_3 - X_3Y_1)^2 + (\Sigma X_i^2)(\Sigma Y_i^2) - (\Sigma X_iY_i)^2 \geq 0.$$

Because each term in the RHS of equation (7.1) is perfect squares, it implies that LHS ≥ 0.

$$(\Sigma X^2)(\Sigma Y^2) - (\Sigma XY)^2 \geq 0 \tag{7.2}$$

$$1 - r^2 = 1 - \left[\frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}\right]^2 = \frac{\sum X^2 \sum Y^2 - \left[\sum XY^2\right]}{\sum X^2 \sum Y^2} \tag{7.3}$$

using (7.2) in (7.3), we have $[1 - r^2] \geq 0$; $r^2 \leq 1$

$r \leq +1$ and $r \leq -1$; implies that $-1 \leq r \leq 1$

Hence, the correlation coefficient lies in the closed interval (–1, 1).

7.5 Probable Error of the Coefficient of Correlation

Normally, we use sample data to evaluate correlation coefficient. So, whenever the result is interpreted, it is necessary to check the reliability of the evaluated sample correlation with the population's coefficient. This is determined by probable error. It is evaluated using the result.

$$\text{Probable error (PE)} = 0.6745 \times (\text{Standard Error [SE] of } r)$$

where SE of $r = \frac{1-r^2}{\sqrt{n}}$

PE of $r = 0.6745 \times \frac{1-r^2}{\sqrt{n}}$

where r is the correlation coefficient, and n is the number of pairs of items. The interpretation is that if PE of $r = \pm\alpha$, where α is a constant, then the range of the correlation of the population can be evaluated approximately as $[r - \alpha, r + \alpha]$.

This PE calculation can be used only when the whole data is normal or nearly normal. The selection of sample should be unbiased. In relation to the PE, the significance of the coefficient of correlation may be judged as follows:

The coefficient of correlation is significant if it is more than 6 times the PE, or when the PE is not much, r exceeds 0.5. It is not significant at all if it is less than the PE.

Example:

Calculate the correlation coefficient and its PE from the following results:

$n = 12$	$\Sigma(X - \overline{X})^2 = 360$	$\Sigma(Y - \overline{Y})^2 = 250$	$\Sigma(X - \overline{X})(Y - \overline{Y})$ $= 225$

And find its PE.

Given,

$n = 12$	$\Sigma(X - \overline{X})^2 = 360$	$\Sigma(Y - \overline{Y})^2 = 250$	$\Sigma(X - \overline{X})(Y - \overline{Y}) = 225$

By definition,

$$r = \frac{\sum_{i=1}^{n}[X_i - \bar{X}][Y_i - \bar{Y}]}{\sqrt{\sum_{i=1}^{n}[X_i - \bar{X}]^2 \sum_{i=1}^{n}[Y_i - \bar{Y}]^2}} = \frac{225}{\sqrt{360 \times 250}} = 0.75$$

$$\text{PE of } r = 0.6745 \times \frac{1-\gamma^2}{\sqrt{n}} = [0.6745 \times 1 - (0.75)^2]/\sqrt{12} = 0.0851$$

The correlation coefficient, 0.75, implies that Y is positively related. The PE of r is 0.0851.

Example:

Calculate coefficient of correlation between X and Y.

	***X* Series**	***Y* Series**
No. of items	15	15
Arithmetic mean	25	18
Squares of deviation from mean	136	138

The sum of the product of deviations X and Y series from their respective means is 122.

Given,

***X* Series**	***Y* Series**
$n_1 = 15$	$n_2 = 15$
$\bar{X} = 25$	$\bar{Y} = 18$

$$\Sigma(X - \bar{X})^2 = 136;\ \Sigma(Y - \bar{Y})^2 = 138;\ \text{and } \Sigma(X - \bar{X})(Y - \bar{Y}) = 122.$$

By definition,

$$r = \frac{\sum_{i=1}^{n}[X_i - \bar{X}][Y_i - \bar{Y}]}{\sqrt{\sum_{i=1}^{n}[X_i - \bar{X}]^2 \sum_{i=1}^{n}[Y_i - \bar{Y}]^2}} = \frac{122}{\sqrt{136 \times 138}} = 0.89$$

The relationship between the variables is positive.

Example:

Evaluate the correlation coefficient for the following data:

$$\Sigma X = 24;\ \Sigma Y = 44;\ n = 4;\ \Sigma X^2 = 164;\ \Sigma Y^2 = 574;\ \text{and } \Sigma XY = 306$$

Consider the given data

$$\Sigma X = 24; \Sigma Y = 44; n = 4; \Sigma X^2 = 164; \Sigma Y^2 = 574; \text{ and } \Sigma XY = 306$$

By definition,

$$r = \frac{n\sum XY - \left[\sum X\right]\left[\sum Y\right]}{\sqrt{\left(\left[n\sum X^2\right]-\left[\sum X\right]^2\right)\left(\left[n\sum Y^2\right]-\left[\sum Y\right]^2\right)}}$$

$$r = \frac{(4\times 306) \quad (24\times 44)}{\sqrt{(4\times 164-(24)^2)\times(4\times 574-(44)^2)}} = \frac{168}{\sqrt{(80)\times(360)}} = \frac{168}{169.71} = 0.99$$

The variables are positively related.

7.6 Rank Correlation

Pearson's correlation coefficient *r* provides a numerical measure of the degree of relationship that exists between the two variables, *X* and *Y*. Also, it requires that the joint distribution of *X* and *Y* be normal. These 2 conditions are not necessary for the rank correlation coefficient. It is based on the ranking of the variates. This was introduced by Charles Edward Spearman in 1904. It helps in dealing with qualitative characteristics like beauty, intelligence, and such items. It is more suitable if the variables can be arranged in order of merit. This is denoted by *R*.

Consider *n* pairs (X_1,Y_1), (X_2,Y_2), …, (X_n,Y_n).

Rank the elements of *X*-series by comparing each and every element of it.

Let it be $R_1, R_2, \ldots, R_n$.

Similarly for *Y* series, let the rank be $S_1, S_2, \ldots, S_n$.

$$\bar{R} = \frac{\sum_{i=1}^{n} R_i}{n} = \frac{1+2+3+\ldots+n}{n} = \frac{n\,[n+1]}{n} = n+1;\ \bar{R} = n+1,$$

Similarly proceeding, we have $\bar{S} = n+1$.

$$\sigma_R{}^2 = \left(\frac{\sum_{i=1}^{n} R_i{}^2}{n}\right) - \left[\frac{\sum_{i=1}^{n} R_i}{n}\right]^2$$

$$\sigma_R{}^2 = \frac{n\times[n+1]\times[2\times n+1]}{6\times n} - \left(\frac{n\times[n+1]}{n\times 2}\right)^2$$

$$\sigma_R{}^2 = \frac{n^2 - 1}{12}$$

Similarly proceeding, we have $\sigma_S{}^2 = \frac{n^2-1}{12}$

$$\text{If } d_i = R_i - S_i\text{; for all } i = 1, \ldots, n;\ d_i = [R_i - \overline{R}] - [S_i - \overline{S}];$$

$$d_i^2 = \{[R_i - \overline{R}] - [S_i - \overline{S}]\} = [R_i - \overline{R}]^2 + [S_i - \overline{S}]^2 - 2\,[R_i - \overline{R}] \times [S_i - \overline{S}]$$

$$\Sigma d_i^2 = \Sigma[R_i - \overline{R}]^2 + \Sigma[S_i - \overline{S}]^2 - 2\,\Sigma\,\{[R_i - \overline{R}]\,[S_i - \overline{S}]\}$$

$$\Sigma d_i^2 = n\sigma_R{}^2 + n\sigma_S{}^2 - 2\Sigma\,\{[R_i - \overline{R}]\,[S_i - \overline{S}]\} = 2 \times \frac{n^2 - 1}{12} - 2\,\Sigma\,\{[R_i - \overline{R}]\,[S_i - \overline{S}]\}$$

$$\Sigma\,\{[R_i - \overline{R}]\,[S_i - \overline{S}]\} = \frac{n^2 - 1}{12} - \frac{1}{2}\sum d_i{}^2$$

$$\text{By definition, } R = \frac{\sum [R_i - \overline{R}][S_i - \overline{S}\ ]}{n * \sigma_R \sigma_S};\ R = \frac{\left[\dfrac{n^2 - 1}{12} - \dfrac{1}{2}\sum R_i{}^2\right]}{\left(\dfrac{n^2 - 1}{12}\right)}$$

$$R = 1 - \frac{6 \times \sum_{i=1}^{n} R_i{}^2}{n \times [n^2 - 1]}$$

Note for Repeated Ranks

The given formula holds good, if the ranks are not repeated. For repeated ranks, say if a rank is repeated for *m* number of times, then the value $\{[m(m - 1)^2]/12\}$ should be added along with (Σdi^2). This must be carried over for each repeated ranks.

Advantages of Rank Correlation Coefficient

1. It is simple to understand and easy to evaluate.
2. It is useful for the qualitative type of data.
3. It can be used even for a quantitative type of data.

Example:

The ranking of 10 trainees in 2 skills, programming and analysis, is as follows. What is the coefficient of rank correlation?

Programming	3	5	8	4	7	10	2	1	6	9
Analysis	6	4	9	8	1	2	3	10	5	7

Because the given data set contains ranks, evaluate the difference in ranks.

Programming (x)	**Analysis (y)**	$R = x - y$	R^2
3	6	−3	9
5	4	1	1
8	9	−1	1
4	8	−4	16
7	1	6	36
10	2	8	64
2	3	−1	1
1	10	−9	81
6	5	1	1
9	7	2	4
			214

$n = 10$. By definition,

$$R = 1 - \frac{6 \times \sum_{i=1}^{n} R_i^2}{n \times [n^2 - 1]} = 1 - \frac{6 \times 214}{10 \times [10^2 - 1]} = -0.2969 = -0.3$$

The rank correlation coefficient is negative, it implies that the variables are negatively related.

Example:

Competitors in a beauty contest are ranked by 3 judges in the following order:

Judge 1 (J_1)	1	6	5	10	3	2	4	9	7	8
Judge 2 (J_2)	3	5	8	4	7	10	2	1	6	9
Judge 3 (J_3)	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to determine which pair has the nearest approach to common taste in beauty.

Because the data set contains ranks, first evaluate the rank correlation coefficient between (J_1, J_2), (J_2, J_3), and (J_3, J_1).

			$[R_{12}]$					
J_1	J_2	J_3	$J_1 - J_2$	D_{12}^2	$R_{23} J_2 - J_3$	D_{23}^2	$R_{31} J_1 - J_3$	D_{31}^2
1	3	6	−2	4	−3	9	−5	25
6	5	4	1	1	1	1	2	4
5	8	9	−3	9	−1	1	−4	16
10	4	8	6	36	−4	16	2	4
3	7	1	−4	16	6	36	2	4
2	10	2	−8	64	8	64	0	0
4	2	3	2	4	−1	1	1	1
9	1	10	8	64	−9	81	−1	1
7	6	5	1	1	1	1	2	4
8	9	7	−1	1	2	4	1	1
				200		214		60

$\Sigma D_{12}{}^2 = 200$; $\Sigma D_{23}{}^2 = 214$; $\Sigma D_{31}{}^2 = 60$; $n = 10$

$$R_{12} = 1 - \frac{6 \times \sum_{i=1}^{n} D_{12}^2}{n \times [n^2 - 1]} 1 - \frac{6 \times 200}{10 \times [10^2 - 1]} = -0.21$$

$$R_{23} = 1 - \frac{6 \times \sum_{i=1}^{n} D_{23}^2}{n \times [n^2 - 1]} = 1 - \frac{6 \times 214}{10 \times [10^2 - 1]} = -0.30$$

$$R_{31} = 1 - \frac{6 \times \sum_{i=1}^{n} D_{31}^2}{n \times [n^2 - 1]} = 1 - \frac{6 \times 60}{10 \times [10^2 - 1]} = 0.6363$$

Judges 1 and 3 have the nearest approach to common taste in beauty.

Example:

Students got the following marks in economics and statistics, respectively. Calculate the rank correlation coefficient.

Marks in Statistics	8	62	36	65	98	39	25	75	82	92
Marks in Economics	84	58	51	35	91	49	60	68	62	86

Rank the given set of data.

Consider the series 1, choose the highest value and give as rank 1, choose the next highest give 2, and proceed till all the entries of the series are ranked. Then repeat the same for series 2.

Marks in Statistics (X)	**Marks in Economics (Y)**	**Rank (X), x**	**Rank (Y), y**	$R = x - y$	R^2
8	84	10	3	7	49
62	58	6	7	–1	1
36	51	8	8	0	0
65	35	5	10	–5	25
98	91	1	1	0	0
39	49	7	9	–2	4
25	60	9	6	3	9
75	68	4	4	0	0
82	62	3	5	–2	4
92	86	2	2	0	0
					92

Here $n = 10$. By definition, the rank correlation

$$r = 1 - \frac{6 \times \sum_{i=1}^{n} R_i{}^2}{n \times [n^2 - 1]} = 1 - \frac{6 \times 92}{10 \times [10^2 - 1]} = 0.44$$

The variables are positively related.

Example:

Find the rank correlation coefficient of the following data.

Series A	115	109	112	87	98	120	98	100	98	118
Series B	75	73	85	70	76	82	65	73	68	80

Consider the data given and rank it.

Series A:

98 repeated for 3 times, so the corresponding rank positions are 7, 8, and 9.

Rank(98) = (7 + 8 + 9)/3 = 8.

A	B	Rank(A), x	Rank(B), y	$R = x - y$	R^2
115	75	3	5	−2	4
109	73	5	6.5	−1.5	2.25
112	85	4	1	3	9
87	70	10	8	2	4
98	76	8	4	4	16
120	82	1	2	−1	1
98	65	8	10	−2	4
100	73	6	6.5	−0.5	0.25
98	68	8	9	−1	1
118	80	2	3	−1	1
					42.5

Series B:

73 is repeated for 2 times, so the corresponding rank positions are 6 and 7.

Rank(73) = (6 + 7)/2 = 6.5

As per Spearman's modified formula for repeated values, along with Σd^2, for each repeated value the element [{$m \times (m^2 - 1)$}/12] should be added. Where m is the number of times the value is repeated.

$$\text{Hence } R = 1 - \frac{6 \times \left\{ \sum_{I=1}^{N} R_i^2 + T \right\}}{n \times [n^2 - 1]}$$

$$\Sigma d^2 = 42.5;\ n = 10;\ T = 2.5;\ R = 1 - \frac{6 \times \{42.5 + 2.5\}}{10 \times [10^2 - 1]} = 0.7272 = 0.73$$

Series	Repeated Value	No of Times (m)	$m \times (m^2 - 1)/12$
A	98	3	$3(9-1)/12 = 2$
B	73	2	$2(4-1)/12 = 1/2$
			$T = 2.5$

The variables are positively related.

Example:

The coefficient of rank correlation between marks in mathematics and statistics of a class is 9/11, and the sum of the squares of the differences in ranks is 30. Find the number of students in the class.

Given $R = 9/11$ and $\Sigma d^2 = 30$.

Find the value of n.

By definition,

$$R = 1 - \frac{6 \times \left\{ \sum_{I=1}^{N} R_i^{\,2} + T \right\}}{n \times [n^2 - 1]} \tag{7.4}$$

Using the given values in the relation (7.4),

$$\frac{9}{11} = 1 - \frac{6 \times [30]}{n \times [n^2 - 1]}; \frac{6 \times [30]}{n \times [n^2 - 1]} = 1 - \frac{9}{11}; n\left(n^2 - 1\right) = 90 \times 11$$

$(n - 1)(n)(n + 1) = 990 = 9 \times 10 \times 11$

Comparing the values of the factors or both LHS and RHS, it is found that $n = 10$.

Hence, the number of students in the class is 10.

7.7 Regression Equations

Regression The word ‘regression’ was first used by Sir Francis Galton in his investigation regarding heredity. Regression means stepping back. This term ‘regression’ is not used in this sense now in statistics. It is a mathematical measure that refers the relationship between 2 variables. This is used to predict the expected value of 1 variable if the value of another is given. Among the 2 variables, one should be treated as an independent variable and the other as dependent.

This relationship can be expressed in the form of a linear equation in 2 variables. Among the 2 variables, X and Y, at a time can be treated as dependent on the other.

(1) X depends on Y (2) Y depends on X.

Regression equation Y depends on X

Consider n pairs of data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ and let the linear equation representing these n data be

$$Y = aX + b \tag{7.5}$$

Take summation on either side of (7.5), $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} [aX_i + b]$

$$\sum_{i=1}^{n} Y_i = a\sum_{i=1}^{n} X_i + nb \tag{7.6}$$

Multiply both sides of the equation (7.5) by X.

$$XY = aX^2 + bX \tag{7.7}$$

Take summation on either side of (7.7),

$$\sum_{i=1}^{n} X_iY_i = \sum_{i=1}^{n}[aX_i^2 + bX_i] = a\sum_{i=1}^{n}[X_i^2] + b\sum_{i=1}^{n}[X_i] \tag{7.8}$$

Equations (7.6) and (7.8) are two linear equations with two unknowns a and b.

Dividing Equation (7.6) by n on both sides, we have

$$\frac{\sum_{i=1}^{n} Y_i}{n} = a \times \frac{\sum_{i=1}^{n} X_i}{n} + b$$

$$\overline{Y} = a \times \overline{X} + b \tag{7.9}$$

Solving the equations (7.5) and (7.9), we have

$$Y - \overline{Y} = a \times [X - \overline{X}] \tag{7.10}$$

n × [7.8] – (ΣX) × [7.6], implies that

$$a = \frac{n\sum_{i=1}^{n} X_iY_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \frac{Cov(X,Y)}{\sigma_X{}^2} \tag{7.11}$$

$$\text{By definition, } r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{7.12}$$

Comparing (7.11) and (7.12), we have

$$a = \frac{r\sigma_Y}{\sigma_X}$$

using the value of a in equation (7.10),

$$Y - \overline{Y} = \frac{r\sigma_Y}{\sigma_X}[X - \overline{X}] \tag{7.13}$$

Equation (7.13) is the required regression equation Y on X.

It is used to estimate the most likely values of Y when the X value is known.

Here, the value $\frac{r\sigma_Y}{\sigma_X}$ is called the 'regression coefficient' of the regression equation Y on X and can be denoted by b_{YX}. Then the equation (7.13) can be expressed as

$$Y - \overline{Y} = b_{YX}[X - \overline{X}]$$

Proceeding in the same way, we can get the regression Equation X depends on Y as

$$X - \overline{X} = \frac{r\sigma_X}{\sigma_Y}[Y - \overline{Y}] \tag{7.14}$$

The value $\frac{r\sigma_Y}{\sigma_X}$ is called the 'regression coefficient' of the regression equation X on Y and can be denoted by b_{XY}. Then the equation (7.14) can be expressed as $X - \overline{X} = b_{XY}[Y - \overline{Y}]$

The equations (7.13) and (7.14) are the required 2 regression equations.

Multiplying the like sides of $b_{XY} = \frac{\gamma\sigma_X}{\sigma_Y}$ and $b_{YX} = \frac{\gamma\sigma_Y}{\sigma_X}$, we have

$$b_{XY} \times b_{YX} = \frac{r\sigma_X}{\sigma_Y} \times \frac{r\sigma_Y}{\sigma_X} = r^2$$

This implies that, $r = \sqrt{b_{XY} \times b_{YX}}$

No.	Nature of b_{XY}	Nature of b_{YX}	Outcome	Nature of r	Nature of Covariance
1.	+	+	+	+	+
2.	–	–	+	–	–
3.	+	–	Not possible		
4.	–	+	Not possible		

NOTE:

- The value of the variances of $\sigma_X{}^2$ and $\sigma_Y{}^2$ is always positive.
- The two regression equations (7.13) and (7.14) imply that the 2 lines are passing through the common point $[\overline{X},\overline{Y}]$.
- To get the value of the 2 means, it is sufficient to solve the given 2 regression equations.

Example:

You are given the data relating to purchase and sales. Obtain 2 regression equations by the method of least squares and estimate the likely sales when the purchases are equal to 100.

Purchases	62	72	98	76	81	56	76	92	88	49
Sales:	112	124	131	117	132	96	120	136	97	85

Let X and Y be the 2 random variables stand for purchases and sales, respectively. Evaluate the necessary summations using the given data.

Purchases (*X*)	**Sales (*Y*)**	***x* = *X* – *a*; *a* = 81**	***y* = *Y* – *b*; *b* = 132**	***xy***	***x*²**	***y*²**
62	112	–19	–20	380	361	400
72	124	–9	–8	72	81	64
98	131	17	–1	–17	289	1
76	117	–5	–15	75	25	225
81	132	0	0	0	0	0
56	96	–25	–36	900	625	1296
76	120	–5	–12	60	25	144
92	136	11	4	44	121	16
88	97	7	–35	–245	49	1225
49	85	–32	–47	1504	1024	2209
Total		–60	–170	2773	2600	5580

By definition,

$$b_{YX} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \frac{(10\times 2773) - (-60)(-70)}{(10\times 2600) - (-60)^2} = 0.783$$

Similarly,

$$b_{XY} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} Y^2{}_i - \left(\sum_{i=1}^{n} Y_i\right)^2} = \frac{(10\times 2773) - (-60)(-170)}{(10\times 5580) - (-170)^2} = 0.652$$

$$\overline{X} = a + [1/n] \times \Sigma X = a + (-60/10) = 81 - 6 = 75 \text{ and}$$

$$\overline{Y} = b + [1/n] \times \Sigma Y = b + (-170/10) = 132 - 17 = 115$$

The regression equation *Y* on *X* is

$$Y - \overline{Y} = b_{YX}\,(X - \overline{X});\ Y - 115 = 0.783\,(X - 75);\ Y - 115 = 0.783X - 58.725$$

$$Y = 0.783X + 56.275 \tag{7.15}$$

The regression equation *X* on *Y* is

$$X - \overline{X} = b_{XY}\,(Y - \overline{Y});\ X - 75 = 0.652\,(Y - 115);\ X - 75 = 0.652Y - 74.98$$

$$X = 0.652Y + 0.02 \tag{7.16}$$

To find the sales when purchase equals 100, put *X* = 100 in equation (7.15),

$$Y = 0.783 \times 100 + 56.275 = 134.575$$

Hence, the amount of sales is 134.575 when the purchase equals 100.

Alternate method

Purchases (X)	Sales (Y)	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$
62	112	–13	–3	39	169	9
72	124	–3	9	–27	9	81
98	131	23	16	368	529	256
76	117	1	2	2	1	4
81	132	6	17	102	36	289
56	96	–19	–19	361	361	361
76	120	1	5	5	1	25
92	136	17	21	357	289	441
88	97	13	–18	–234	169	324
49	85	–26	–30	780	676	900
750	1150			1753	2240	2690

$$\bar{X} = \Sigma X/n = 750/10 = 75;\ \bar{Y} = \Sigma Y/n = 1150/10 = 115$$

$$b_{yx} = \frac{\sum_{i=1}^{n}[X_i - \bar{X}][Y_i - \bar{Y}]}{\sum_{i=1}^{n}[X_i - \bar{X}]^2} = 1753/2240 = 0.783$$

$$b_{xy} = \frac{\sum_{i=1}^{n}[X_i - \bar{X}][Y_i - \bar{Y}]}{\sum_{i=1}^{n}[Y_i - \bar{Y}]^2} = 17530/26900 = 0.652$$

Using the values of $\bar{X}$, $\bar{Y}$, *bxy,* and *byx,* you can construct the 2 regression equations as stated.

Example:

The following table gives aptitude test scores and productivity indices of 8 randomly selected workers: Find the equation to the line that can be used to predict the productivity index from the aptitude score. Estimate the productivity index of a worker whose test score is 66.

Aptitude score (X)	57	58	59	59	60	61	62	63
Productivity index (Y)	67	68	65	68	72	72	69	71

Evaluate the necessary summations using the given data.

Aptitude Score (x)	Productivity Index (y)	$X = x - 60$	$Y = y - 68$	XY	X^2	Y^2
57	67	–3	–1	3	9	1
58	68	–2	0	0	4	0
59	65	–1	–3	3	1	9
59	68	–1	0	0	1	0
60	72	0	4	0	0	16
61	72	1	4	4	1	16
62	69	2	1	2	4	1
63	71	3	3	9	9	9
		–1	8	21	29	52

By definition, $X = x - 60$ and $Y = y - 68$

$$b_{yx} = \frac{n\sum_{i=1}^{n} X_iY_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \frac{8(21) - (-1)(8)}{8(29) - (-1)^2} = 0.762$$

$$b_{xy} = \frac{n\sum_{i=1}^{n} X_iY_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2} = \frac{8(21) - (-1)8}{8(52) - (8)^2} = 0.5$$

$$\bar{x} = a + [1/n] \times \sum X = 60 + (-1/8) = 60 - 0.125 = 59.875 \text{ and}$$

$$\bar{y} = b + [1/n] \times \sum Y = 68 + (8/8) = 68 + 1 = 69$$

The regression equations are

$$y - \bar{y} = b_{yx}\,(x - \bar{x});\; y - 69 = 0.762 \times (x - 59.875);\; y - 69 = 0.762 \times x - 45.625$$

$$y = 0.762 \times x + 23.375$$

$$x - \bar{x} = b_{xy}(y - \bar{y});\; x - 59.875 = 0.5 \times (y - 69);\; x = 0.5\,y + 25.375$$

Hence, the required 2 regression equations are

$$y = 0.762 \times x + 23.375 \tag{7.17}$$

$$x = 0.5\,y + 25.375 \tag{7.18}$$

Additionally, we have to evaluate the productivity index (y) when the aptitude scores (x) is 66. Put $x = 66$ in (7.17), we have $y = 0.762 \times 66 + 23.375$; $Y = 73.667$

Hence, the employee's test score is 66 and the productivity index is 73.667.

Example:

The correlation coefficient between supply (Y) and price (X) of a commodity is 0.60. If $\sigma_X = 150$, $\sigma_Y = 200$, mean $[X] = 10$ and mean $[Y] = 20$. Find the equations of the regression lines of Y on X and X on Y.

Given $Y = 0.6$; $\sigma_X = 150$, $\sigma_Y = 200$, mean $[X] = 10$ and mean $[Y] = 20$.

By definition,

$$b_{XY} = \frac{\gamma\sigma_X}{\sigma_Y} = \frac{0.6 \times 150}{200} = 0.45$$

$$b_{YX} = \frac{\gamma\sigma_Y}{\sigma_X} = \frac{0.6 \times 200}{150} = 0.8$$

The regression equation Y on X is $Y - \overline{Y} = b_{YX}(X - \overline{X})$

$$Y - 20 = 0.8(X - 10) = 0.8X - 8;\ Y = 0.8 \times X + 12 \tag{7.19}$$

The regression equation X on Y is

$$X - \overline{X} = b_{XY}(Y - \overline{Y});\ X - 10 = 0.45(Y - 20);\ X = 0.45Y + 1 \tag{7.20}$$

The regression equation Y on X is $Y = 0.8X + 12$.

The regression equation X on Y is $X = 0.45Y + 1$.

Example:

In a partially destroyed laboratory record of an analysis of correlated data, the following results only are legible:

Regression equations: $8X - 10Y + 66 = 0$; $40X - 18Y = 214$.

What were (1) the mean values of X and Y?

(2) The correlation coefficient between X and Y?

(3) If $\sigma_X{}^2 = 9$, find the value of σ_Y?

Consider the two regression equations,

$$8X - 10Y + 66 = 0 \tag{7.21}$$

$$40X - 18Y = 214 \tag{7.22}$$

We must choose an equation for X on Y and the other for Y on X.

Because the magnitude of coefficient of Y in Equation (7.21) is dominating the magnitude coefficient of X, choose Equation (7.21) for Y on X, and Equation (7.22) for X on Y.

$$\text{Equation (7.21) can be rewritten as, } 10Y = 8X + 66\text{, then } Y = 0.8X + 6.6 \tag{7.23}$$

$$\text{Equation (7.22) can be rewritten as, } 40X = 18Y + 214\text{, then } X = 0.45Y + 5.35 \tag{7.24}$$

Comparing Equation (7.23) with the actual equation $Y = b_{YX} \times X + C_1$

we have, $b_{YX} = 0.8$

In the same way, comparing Equation (7.24) with the actual equation

$X = b_{XY} \times Y + C_2$; we have, $b_{XY} = 0.45$

$$\text{By definition, } b_{XY} = \frac{\gamma\sigma_X}{\sigma_Y} = 0.8 \tag{7.25}$$

$$\text{and } b_{xy} = \frac{\gamma\sigma_Y}{\sigma_X} = 0.45 \tag{7.26}$$

Multiplying the like sides of Equations (7.25) and (7.26) we have,

$$r^2 = 0.8 \times 0.45 = 0.36;\ r = \pm\ 0.6$$

Because both the regression coefficients are positive, the value of the correlation coefficient must be positive.

Hence, the value of correlation coefficient is 0.6. To get the mean values of X and Y solve the 2 given equations (7.27) and (7.28) for X and Y. The value of X is taken to be the mean value of X and the value of Y is taken to be the mean value of Y.

$$8X - 10Y + 66 = 0 \tag{7.27}$$

and

$$40X - 18Y = 214 \tag{7.28}$$

$$5 \times [1] - [2]\text{, implies that } -32Y = -544;\ Y = 17.$$

Using the values of $Y = 17$ in Equation (7.27) we have $X = 13$.

Hence, the mean of X is 13 and the mean of Y is 17.

Given $\sigma_X{}^2 = 9$; using the value of σ_X and Y in Equation (7.25),

$$0.6 \times [\sigma_Y/3] = 0.8;\ \sigma_Y = 4.$$

NOTE: In a special case, if the dominancy among the coefficients of the variables are not existing (in both the equation coefficient X dominates or Y dominates), choose any of the equations for Y on X and the other one for X on Y based on trial and error. This selection should satisfy the condition $b_{YX} \times b_{XY} \leq 1$. If this condition fails, then revert the selection and proceed.

Example:

Two lines of regressions are given by $x + 2y = 5$ and $2x + 3y = 8$. Calculate the value of mean of X and mean of Y and Υ.

Consider the given regression equations,

$$x + 2y = 5 \tag{7.29}$$

and

$$2x + 3y = 8 \tag{7.30}$$

There is no pure dominance existing among the 2 variables in both the equations. Clearly the coefficient of Y dominates in terms of magnitude in both the equations. Choosing equation (7.29) for Y on X based on trial and error,

$$\text{Rewriting } x + 2y = 5 \text{ as } 2y = -x + 5; y = (-1/2)x + (5/2) \tag{7.31}$$

Equation (7.31) implies that $b_{yx} = -0.5$

Choose the second equation for X on Y.

$$\text{Rewriting } 2x + 3y = 8; 2x = -3y + 8; x = (-3/2)y + 4 \tag{7.32}$$

Then we have, $b_{xy} = -1.5$

$$b_{xy} \times b_{yx} = (-3/2)(-1/2) = 3/4 \leq 1$$

Hence, the selection is correct. If *bxy* × *byx* >1; then change the selection of equation for Y on X and X on Y then proceed.

$$\text{By definition, } b_{XY} = \frac{\gamma\sigma_X}{\sigma_Y} = -0.5 \tag{7.33}$$

$$\text{and } b_{YX} = \frac{\gamma\sigma_Y}{\sigma_X} = -1.5 \tag{7.34}$$

Multiplying the like sides of Equations (7.33) and (7.34) we have,

$$r^2 = [-0.5] \times [-1.5] = 0.75; r = \pm 0.866$$

Because both the regression coefficients are negative, the value of the correlation coefficient must be negative.

Hence, the value of correlation coefficient is −0.866

To get the mean values of x and y, solve Equations (7.29) and (7.30) for x and y. The value of x is taken to be the mean value of x, and the value of y is taken to be the mean value of y.

Solving the Equations (7.29) and (7.30), we have $x = 1$ and $y = 2$.

Hence, the mean of $x = 1$ and the mean of $y = 2$.

Exercise 7

1. A study by a roadway company on the effect of bus-ticket prices on the number of passengers produced the following results.

Ticket Price ($)	Passengers/100 km
25	800
30	780
35	780
40	660
45	640
50	600
55	620
60	620

Develop the estimating equation that best describes the data. Predict the number of passengers per 100 miles if the ticket price is $10.

2. Calculate the coefficient of correlation between the age group and the rate of mortality from the following data:

Age group	0–20	20–40	40–60	60–80	80–100
Rate of mortality	350	280	540	760	900

3. The following table gives the distribution of production and also the relatively defective item among them, according to size groups. Is there any correlation between group size and number of defective items?

Size Group	15–16	16–17	17–18	18–19	19–20	20–21
No. of items	200	270	340	360	400	300
No. of defectives	150	162	170	180	180	120

4. Find the Spearman's correlation coefficient after making adjustments for tied ranks.

X	91	88	86	85	76	70	62	85	52	50
Y	85	82	90	76	84	51	81	67	36	56

5. Competitors in a beauty contest are ranked by 3 judges. Find which pair of judges has the nearest approach to common taste in beauty:

I	1	6	5	10	3	2	4	9	7	8
II	3	5	8	4	7	10	2	1	6	9
III	6	4	9	8	1	2	3	10	5	7

6. Given the regression lines as $3x + 2y = 26$ and $6x + y = 31$. Find their point of intersection and interpret it. Also find the correlation coefficient between x and y.

7. Find the most likely price in Mumbai corresponding to the price \$70/– at Kolkata from the following:

	Kolkatta	**Mumbai**
Average price	65	67
Standard deviation	2.5	3.5

The correlation coefficient between the prices of commodities in the two cities is 0.8.

8. The mean sales of a firm is \$40 lakhs with SD \$10 lakhs. The average advertisement expenditure of that firm is \$6 lakhs with SD 1.5 lakhs. The coefficient of correlation between sales and the advertisement expenditure is 0.9. (a) Estimate the likely sales for a proposed advertisement expenditure of \$10 lakhs. (b) What should the advertisement budget be, if the company wants to attain a sales target of \$60 lakhs.

9. The following results of the capital employed and profit earned by a firm in 10 successive years are given:

	Mean	**SD**
Capital employed (\$ in thousands))	55	28.7
Profit earned (\$ in thousands)	13	8.5

Coefficient of correlation is 0.96. (a) Obtain the 2 regression lines. (b) Estimate the amount of profit to be earned if capital employed is \$50000. (c) Estimate the amount of capital to be earned if the profit earned is \$20000.

10. Evaluate the value of r for the following data:

Sales Revenue (Lakhs of \$)	**Advertisement Expenditure (\$ in thousands)**			
	5–15	**15–25**	**25–35**	**35–45**
75–125	3	4	4	8
125–175	8	6	5	7
175–225	2	2	3	4
225–275	2	3	2	2

11. Find the value of r between the profits and sales for the following data:

	Sales ($ in thousands)				
Profits ($ in thousands)	**80–90**	**90–100**	**100–110**	**110–120**	**120–130**
50–55	1	3	7	5	2
55–60	2	4	10	7	4
60–65	1	5	12	10	7
65–70	—	3	8	6	3

12. An operation analyst conducts a study to analyze the relationship between production, X, and manufacturing expenses, Y, in the electronics industry. A sample of $n = 10$ firms, randomly selected from within the industry, yields the data in the following table; manufacturing expenses are considered to be independent variable. They changes as the volume of production varies. On the other hand, a change in manufacturing expenses would not necessarily cause a change in the volume of production.

X (units in thousands)	40	42	48	55	65	79	88	100	120	140
Y ($ in thousands)	150	140	160	170	150	162	185	165	190	185

Construct the regression equation y on x.

13. An analyst is studying the relationship between shopping centre traffic and a department store's daily sales. The analyst develops an index to measure the daily volume of traffic entering the shopping centre and an index of daily sales. The following table shows the index values for 10 randomly selected days.

Traffic index, X	71	82	111	85	89	110	111	121	129	132
Sales index, Y	250	280	301	325	328	390	410	420	450	475

14. The manager of a Graduate House West dining facility wants to be able to predict how many extra meals to prepare on days when conferences are held at the university conference centre. A random sample of records for the past 2 years yields the following:

No. of Conference Registrants, X	126	367	213	140	154	186	111	411
No. of Extra Meals, Y	103	287	169	122	137	140	80	328
No. of Conference Registrants, X	321	260	151	149	237	329	159	218
No. of Extra Meals, Y	283	201	118	116	185	289	153	202
No. of Conference Registrants, X	185	178	199	107	301	270	285	345
No. of Extra Meals, Y	174	164	158	79	210	165	265	250

Construct the regression lines, and hence, evaluate the number of extra meals required when the conference registrants are 200.

15. Consider the details of the import and export (units are represented in lakh of rupees) of Bhavana Shree Limited. Construct the 2 regression lines and evaluate the value of export corresponding to the import of 40 lakhs.

Export, X	38.9	29.3	28.3	25.2	22	21.5	17.5	17	14.3
Import, Y	43.4	25	34.7	25.6	25.5	26.8	14.9	22.2	21.6

16. For a sample of 8 employees, a personnel director has collected the following data on ownership of company stock versus years with the firm. Find the value of r and the regression equations.

Years, X	6	12	14	6	9	13	15	9
Shares, Y	300	408	560	252	288	650	630	522

17. The following data represent boat sales (x) and boat trailer sales (y) from 2002 to 2007. Determine the value of r and the regression equations.

Year	Boat Sales ($ in thousands)	Boat Trailer Sales ($ in thousands)
2002	649	207
2003	619	194
2004	596	181
2005	576	174
2006	585	168
2007	574	159

18. The following ratings are based on collision claim experience and their frequency for 12 makes of small, 2-door cars. Higher numbers reflect higher claims and more frequent thefts, respectively. Find the regression equations.

Collision	Theft	Collision	Theft
103	103	106	97
97	113	139	425
108	81	110	82
115	68	96	81
127	90	84	59
104	79	105	167

8

Probability

8.1 Introduction

The concept of probability was introduced late in the seventeenth century. This concept was introduced in problems relating to games of probability (i.e., tossing a coin, playing cards). But the probability concept is now used in almost all areas of study such as economics, statistics, industry, engineering, and business. Probability is related to the study of events that are going to happen or not.

Before going further, let's define some of the basic terms that are going to be used in the definition of probability.

8.2 Definitions for Certain Key Terms

8.2.1 Experiment

An experiment means an activity or measurement that result in an outcome.

> **Example:**
>
> Tossing a single coin for 50 times.

8.2.2 Sample Space

Sample space refers to the collection of all possible events of an experiment, denoted by S.

> **Example:**
>
> In a coin-tossing experiment, the sample space should contain the possible outcomes of a head (H) or a tail (T); S = {H, T}

8.2.3 Event

Event means one or more of the possible outcomes of an experiment; it is a subset of a sample space.

> **Example:**
>
> In throwing a dice, S = {1, 2, 3, 4, 5, 6} contains the face 1 is an event.

8.2.4 Equally Likely Events

In a sample space containing at least 2 events, the chance of the occurrence of each of the event is equal.

Example:

In a coin-tossing experiment, having a head or tail in a trial is equal to ½ each or 50%.

8.2.5 Mutually Exclusive Events

Events are said to be mutually exclusive if the outcome is only 1 element at a time. There is no chance that 2 or more events happen together. Alternatively, it is called an 'incompatible event'.

Example:

In a coin-tossing experiment, we can have either head or tail as an outcome. Clearly the occurrence of head prevents the occurrence of the tail, which implies that the 2 events are said to be mutually exclusive.

8.2.6 Outcome

An outcome is the result of a random experiment.

Example:

In coin tossing, the 2 possible outcomes are head or tail.

8.3 Meaning of Probability

The term 'probability' can be defined in two approaches, the classical approach and relative frequency approach.

8.3.1 The Classical Approach

The classical approach describes the term 'probability' as the proportion of time in the event that can be theoretically expected to happen.

$$\text{Probability} = \frac{\text{Number of possible outcomes in which the event occur}}{\text{Total number of possible outcomes}}$$

Example:

Find the probability of having the face 1 in throwing a dice.

Selection of the face 1: It is one of the outcomes of 6 possible outcomes (equally likely events), that is, 1/6.

8.3.2 The Relative Frequency Approach

In the relative frequency approach, the probability is the proportion of times an event is observed to happen in a large number of trials.

$$\text{Probability} = \frac{\text{Number of trials in which the event occurs}}{\text{Total number of trials}}$$

8.3.3 Notation

The probability of an event A is denoted by P(A). The value of P(A) should be in the range $0 \leq P(A) \leq 1$.

If the event A′ is the negation of the event A, then its probability can be defined as P(A′). Clearly the range of P(A′) is $0 \leq P(A') \leq 1$.

1) This implies that, $P(A) + P(A') = 1$. Also, $P(A) = 1 - P(A')$ and $P(A') = 1 - P(A)$.

NOTE:

- If P(A) = 1, then the event A is said to be a sure event.
- If P(A) = 0, then the event A is said to be a null event.

Example:

If a coin is tossed, what is the chance of a head?

The sample space can be defined as, S = {H, T}; $n(S) = 2$.

Let A be the event refers the occurrence of head, then A = {H}; $n(A) = 1$.

The probability of having head, $P(A) = n(A)/n(S)$.

Here, $n(A)$ is the number of elements in the set A, and $n(S)$ is the number of elements in the set, S. Then P(A) = ½ = 0.5

Example:

Three fair coins are tossed once. Find the probability of

(1) At least 1 tail, (2) Exactly 1 head, (3) Exactly 2 tails, (4) Exactly 3 heads, and (5) At least 2 tails.

The sample space can be defined as,

S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}; $n(S) = 8$.

1. Let A stand for the event of at least one tail to happen, then

   A = {HHT, HTH, HTT, THH, THT, TTH, TTT}; $n(A) = 7$;
   then, $P(A) = n(A)/n(S) = 7/8$.

2. Let B stand for the event of exactly 1 head to happen,

   B = {HTT, THT, TTH}; $n(B) = 3$;
   then, $P(B) = n(B)/n(S) = 3/8$.

3. Let C stand for the event of occurrence of exactly 2 tails,

   C = {HTT, THT, TTH}; n(C) = 3; then P(C) = n(C)/n(S) = 3/8.

4. Let D stand for the event of exactly 3 heads to happen,

   D = {HHH}; n(D) = 1; then P(D) = n(D)/n(S) = 1/8.

5. Let E stand the event of at least 2 tails to occur,

   E = {HTT, TTH, THT, TTT}; n(E) = 4; then P(E) = n(E)/n(S) = 4/8 = ½.

Example:

If a dice is tossed, what is the probability the number appearing on top is (1) odd number, (2) less than 3, and (3) an even number less than 5.

The sample space can be defined as S = {1, 2, 3, 4, 5, 6}; n(S) = 6.

1. Let A be the event of having an odd number, A = {1,3,5}; n(A) = 3; then, P(A) = n(A)/n(S) = 3/6 = 1/2 = 0.5.
2. Let B be the event of having the number less than 3, B = {1, 2}; n(B) = 2; then P(B) = n(B)/n(S) = 2/6 = 1/3 = 0.333
3. Let C be the event of having an even number less than 5, C = {2, 4}; n(C) = 2; then P(C) = n(C)/n(S) = 2/6 = 1/3.

Example:

What is the probability of pulling 2 red balls in a draw of 2 balls from a box containing 4 white and 3 red balls?

Given, Box contains:

3 R Balls	4 W Balls

Number of red balls = 3; Number of white balls = 4

Total number of balls = 7; Number of balls to be selected = 2.

Total number of ways of selecting two red balls out of 7 balls = $^{7}C_{2}$ = (7 × 6)/(1 × 2) = 21.

Number of favourable chances of selecting 2 red balls out of 3 red balls = $^{3}C_{2}$ = (3 × 2)/(1 × 2) = 3.

P(selecting 2 red balls in 2 draws) = 3/21 = 1/7 = 0.143.

Example:

What is the chance that a leap year selected at random will contain 53 Wednesdays?

Number of weeks in a year = 52; Number of days = 52 × 7 = 364

Number of days in a leap year = 366

Difference in days between the leap year and the normal year = 366 − 364 = 2

Clearly, we have 2 excess days. The sample space of the 2 excess days can be given as

S = {(Sun, Mon), (Mon, Tue), (Tue, Wed), (Wed, Thr), (Thr, Fri), (Fri, Sat), (Sat, Sun)}

$$n(S) = 7$$

To get 53 Wednesdays, we must look for the excess of 1 more Wednesday (53 − 52 = 1).

Let A be the event of the occurrence of 53rd Monday.

Then, A = {(Tuesday, Wednesday), (Wednesday, Thursday)}; *n*(A) = 2.

P(having 53 Wednesdays in a leap year) = *n*(A)/*n*(S) = 2/7 = 0.286.

Example:

From a pack of 52 cards, 1 card is drawn at random. Find the chance of drawing a heart and a chance of not drawing a heart.

Total number of cards in a pack = 52

Number of cards to be selected = 1

Total chance of selecting one card out of 52 cards = $^{52}C_1 = 52$

Number of cards having heart symbol = 13

The number of hearts to be selected = 1

A total number of favourable chances = $^{13}C_1 = 13$

Let A be the event of selection of one heart, then P(A) = 13/52 = ¼ = 0.25.

We know that P(A) + P(A′) = 1. P(A′) = 1 − P(A) = 1 − (¼) = 0.75.

The chance of *not* drawing a heart is 0.75.

8.3.4 Addition Rules for Probability

There are situations in which we wish to evaluate the probability that 2 or more of several events will occur in an experiment. The evaluation of such probabilities seeks the help of addition rules.

Events Are Not Mutually Exclusive:

When events are not mutually exclusive, 2 or more of them can happen at the same time. In this case, let us derive the condition based on 2 events.

Result 1

If A and B are any 2 events, then the probability that at least 1 of the 2 events A and B occurs can be given by denoted by P(A∪B) and the same can be defined as

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Let S be the sample space, and A and B be the 2 events of S.

Then by definition,

$$P(A \cup B) = \frac{n(A \cup B)}{n(S)} \tag{8.1}$$

We know that,

$$n(A \cup B) = n(A) + n(B) - n(A \cap B) \tag{8.2}$$

Dividing by n(S) on both sides of Equation (8.2), we have

$$\frac{n(\mathrm{A} \cup \mathrm{B})}{n(\mathrm{S})} = \frac{n(\mathrm{A})}{n(\mathrm{S})} + \frac{n(\mathrm{B})}{n(\mathrm{S})} - \frac{n(\mathrm{A} \cap \mathrm{B})}{n(\mathrm{S})}$$

$$\mathrm{P(A{\cup}B) = P(A) + P(B) - P(A{\cap}B)} \tag{8.3}$$

NOTE: Equation (8.3) can be generalized for any number of events.

Result 2

Let us extend the result of Result 1 for any three events A, B, and C. Find P(A∪B∪C).
Let B∪C = D, then we have P(A∪B∪C) = P(A∪D)

$$\begin{aligned}\mathrm{P(A{\cup}B{\cup}C)} &= \mathrm{P(A) + P(D) - P(A{\cap}D) = P(A) + (B{\cup}C) - P((B{\cup}C){\cap}A)}\\ &= \mathrm{P(A) + P(B) + P(C) - P(B{\cap}C) - P[(B{\cap}A) \cup (C{\cap}A)]}\\ &= \mathrm{P(A) + P(B) + P(C) - P(B{\cap}C) - [P(B{\cap}A) + P(C{\cap}A) - P(A{\cap}B{\cap}C)]}\end{aligned}$$

$$\mathrm{P(A{\cup}B{\cup}C) = P(A) + P(B) + P(C) - P(A{\cap}B) - P(B{\cap}C) - P(A{\cap}C) + P(A{\cap}B{\cap}C)} \tag{8.4}$$

The derived 2 results can be deduced further based on certain conditions on the events.

Condition 1

A, B, and C are 3 mutually exclusive events. When the events are mutually exclusive, then only 1 event can occur at a time. There is no chance for the occurrence of 2 or 3 events together.

The same thing can be expressed as follows:

$$\mathrm{P(A{\cap}B) = P(B{\cap}C) = P(C{\cap}A) = P(A{\cap}B{\cap}C) = 0}$$

Hence the results 1 and 2 can be reduced as follows:

$$\mathrm{P(A{\cup}B) = P(A) + P(B)} \tag{8.5}$$

$$\mathrm{P(A{\cup}B{\cup}C) = P(A) + P(B) + P(C)} \tag{8.6}$$

Condition 2

The events A, B, and C are 3 independent events. When the events are independent, we have,

- $\mathrm{P(A{\cap}B) = P(A) \times P(B)}$
- $\mathrm{P(C{\cap}B) = P(C) \times P(B)}$
- $\mathrm{P(A{\cap}C) = P(A) \times P(C)}$
- $\mathrm{P(A{\cap}B{\cap}C) = P(A) \times P(B) \times P(C)}$

Hence the results 1 and 2 can be reduced as follows:

$$P(A \cup B) = P(A) + P(B) - P(A) \times P(B) \tag{8.7}$$

$$\begin{aligned} P(A \cup B \cup C) = {} & P(A) + P(B) + P(C) - P(A) \times P(B) - P(C) \times P(B) - P(A) \\ & \times P(C) + P(A) \times P(B) \times P(C) \end{aligned} \tag{8.8}$$

Example:

A construction company is bidding for 2 contracts, A and B. The probability that the company will get A is 3/5, the probability that the company will get contract B is 1/3, and the probability that the company will get both the contracts is 1/8. What is the probability that the company will get contract A or B?

Given A and B can be any 2 contracts and the probabilities,

$$P(A) = 3/5;\ P(B) = 1/3;\ P(A \cap B) = 1/8$$

To find, $P(A \cup B)$.

By definition, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$= 3/5 + 1/3 - 1/8 = \frac{72 + 40 \quad 15}{120} = \frac{97}{120} = 0.808$$

The probability of the company to get the contract A or B is 0.808.

Example:

A fair dice is thrown. What is the chance that either an even number or a number greater than 3 will turn up?

The sample space S can be defined as S = {1, 2, 3, 4, 5, 6}; $n(S) = 6$

Let A be the event of having an even number, then A = {2, 4, 6}; $n(A) = 3$

Let B be the event of having a number that is more than 3, then B = {4, 5, 6}; $n(B) = 3$

To find $P(A \cup B)$.

$$A \cap B = \{4, 6\};\ n(A \cap B) = 2$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 3/6 + 3/6 - 2/6 = 4/6 = 2/3 = 0.667$$

The probability of either an even number or a number greater than 3 will turn up is 0.667.

Example:

The probability that a contractor will not get a plumbing contract is 1/3, and the probability that he will get an electrical contract is 4/9. If the probability of getting at least 1 contract is 4/5, what is the probability that he will get both the contracts?

Let A and B stand for the event of getting the plumbing and electrical contracts, respectively.

Given, $P(A') = 1/3$; $P(B) = 4/9$; and $P(A \cup B) = 4/5$. Find $P(A \cap B)$.

$P(A') = 1/3$, implies that $P(A) = 1 - P(A')$; $P(A) = 1 - 1/3 = 2/3$.

By definition, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) = [2/3] + [4/9] - [4/5] = 14/45 = 0.311.$$

The probability that he will get both the contracts is 0.311.

8.3.5 Multiplication Rule on Probability When Events Are Independent

Events Are Independent:

When the occurrence of 1 event has no effect on the probability that another will occur, their joint probability is the product of their individual probabilities, such that then $P(A \cap B) = P(A) \times P(B)$

NOTE: If 2 events A and B are independent, then the following events are also independent

- A' and B
- B' and A
- A' and B'

We have,

- $P(A' \cap B) = P(A') \times P(B)$
- $P(A \cap B') = P(A) \times P(B')$
- $P(A' \cap B') = P(A') \times P(B')$

Example:

A candidate is selected for an interview for 3 posts. In the first post, there are 3 candidates, for the second, there are 4, and for the third, there are 2. What are the chances of his getting at least 1 post?

Let A, B, and C stand for the events of getting selected for Post 1, Post 2 and Post 3, respectively.

No. of candidates for the first post = 3; $P(A) = 1/3 = 0.333$

No. of candidates for the second post = 4; $P(B) = 1/4 = 0.25$

No. of candidates for the third post = 2; $P(C) = 1/2 = 0.5$

To find $P(A \cup B \cup C)$. Here the events A, B, and C are independent.

Let $A \cup B \cup C = D$, then we have, $P(D) + P(D') = 1$.

$$P(D) = 1 - P(D') = 1 - P[(A \cup B \cup C)']$$

Using Demorgon's property, $(A \cup B \cup C)' = A' \cap B' \cap C'$

$$P(A \cup B \cup C)' = 1 - P(A' \cap B' \cap C') = 1 - P(A') \cdot P(B') \cdot P(C')$$

$$= 1 - [1 - (1/3)] \times [1 - (1/4)] \times [1 - (1/2)]$$

$$= 1 - [2/3] \times [3/4] \times [1/2] = 1 - [1/4] = [3/4] = 0.75.$$

The chance of getting at least one post is 0.75.

8.3.6 Compound Probability or Conditional Probability

When events A and B are not independent, the occurrence of A will influence the probability that B will take place. The multiplication rule when A and B are independent can be given as: $P(A \cap B) = P(A) \times P(B/A)$ or $P(B/A) = [P(A \cap B)]/[P(A)]$; where $P(A) > 0$.

Here $P(B/A)$ is the conditional probability referring that the chance of B occurring after the occurrence of A. (The event A occurs first, then followed by the second event, B.)

In the same way, we can define the conditional probability of Event A, given that B has occurred.

$$P(A \cap B) = P(B) \times P(A/B) \text{ or } P(A/B) = [P(A \cap B)]/[P(B)]; \text{ where } P(B) > 0.$$

Example:

The personnel department of a company has records that show the following analysis of its 200 engineers:

Age	Undergraduate Degree Only	Postgraduate Degree Only
<30	90	10
30–40	20	30
>40	40	10

If one engineer is selected at random from the company, find

- The probability that he has only undergraduate (UG) degree.
- The probability that he has postgraduate (PG) degree, given that he is older than 40 years of age.
- The probability that he is younger than 30 years of age, given that he has only a UG degree.

Given,

Age	UG Degree Only	PG Degree Only	Total
<30	90	10	100
30–40	20	30	50
>40	40	10	50
Total	150	50	200

Let A, B, C, and D be the events of selected personnel to have UG degree only, PG degree, with age older than 40, and age younger than 30, respectively.

To find (1) P(A), (2) P(B/C), and (3) P(D/A)

1. $P(A) = \dfrac{\text{Total No. of persons having UG degree only}}{\text{Total employees}}$

From the table, we have P(A) = 150/200 = 0.75; P(A) = 0.75.

2. By definition, P(B/C) = P(C∩B)/P(C)

From the table, we have P(C∩B) = 10/200; and P(C) = 50/200

$$P(B/C) = (10/200)/(50/200) = (10/50) = 0.2$$

3. By definition, P(D/A) = P(D∩A)/P(A)

From the table, we have P(D∩A) = (90/200); P(A) = (150/200)

$$P(D/A) = (90/200)/(150/200) = (9/15) = 0.6,$$

Hence,

- The probability that the engineer has only an UG degree is 0.75
- The probability that the engineer has a PG degree, given that age is older than age 40, is 0.2
- The probability that he is younger than age 30, given that he has only an UG degree is 0.6

Example:

A bag contains 8 red balls and 5 white balls. Two successive draws are made. Find the probability that the first draw will yield 3 white balls and the second 3 red balls.

(1) With replacement and (2) without replacement

Number of red balls = 8

Number of white balls = 5

Total number of balls = 13

1. With replacement:

   First draw: 3 white balls; Total chances = $13C_3$

   Number of favourable chances = $5C_3$

   P(having 3 white balls in the first draw) = $5C_3/13C_3$ = (10/286) = 0.035

   The 3 white balls selected in the first are replaced before the second draw.

   Second draw: 3 red balls; Total chances = $13C_3$

   Number of favourable chances = $8C_3$

   P(second draw/first draw) = $8C_3/13C_3$ = 56/286 = 0.196

   P(3 white balls and the second 3 red balls with replacement) = 0.035 × 0.196 = 0.00686 = 0.0069

2. Without replacement:

   First draw: 3 white balls; P(first draw) = 0.035

   The 4 white balls selected in the first are not replaced before the second draw.

   Second draw: Given that the balls are not replaced.

   Total number of balls after the first draw = 13 – 3 = 10

   Total chances = ${}^{10}C_3$; Number of favourable chance = ${}^{8}C_3$

   P(second draw/first draw) = ${}^{8}C_3/{}^{10}C_3$ = 56/120 = 0.467

   P(3 white balls and the second 3 red balls without replacement) = 0.035 × 0.467 = 0.0163

Example:

Two parties are competing for the position on the Board of Directors of a company. The probabilities that the first and second parties will win the position are 0.65 and 0.35 respectively. If the first party wins, the probability of introducing a new product is 0.75, and the corresponding probability for the second party is 0.4. What is the probability that the new product will be introduced?

Let A and B are the events of Party 1 and Party 2 to win the board of directors of a company, respectively. Let C be the event of introducing the new product.

Given, P(A) = 0.65; P(B) = 0.35; P(C/A) = 0.75 and P(C/B) = 0.4

The event C can happen if

- Party 1 wins and introduces the new product, $(A \cap C)$
- Party 2 wins and introduces the new product, $(B \cap C)$

$C = (A \cap C) \cup (B \cap C)$. Both the events are mutually exclusive.

To find $P(C) = P((A \cap C) \cup (B \cap C)) = P(A \cap C) + P(B \cap C)$; (using addition theorem)

By definition, $P(A \cap C) = P(C/A) \times P(A) = 0.65 \times 0.75 = 0.4875$ and

$$P(B \cap C) = P(C/B) \times P(B) = 0.35 \times 0.4 = 0.14.$$

$$\text{Then, } P(C) = 0.4875 + 0.14 = 0.6275$$

Hence, the chance of introducing the new product is 0.6275.

Example:

There are 3 men ages 60, 65, and 70 years. The probability to live 5 years more is 0.8 for a 60-year-old, 0.6 for a 65-year-old, and 0.3 for a 70-year-old person. Find the probability that at least 2 of the 3 persons will live for another 5 years.

Let A, B, and C be the events of persons with ages 60, 65, and 70 years and that they will live for 5 more years. All the events are independent.

Let D be the event of at least 2 of the 3 persons will live for another 5 years.

Given,

- P(A) = 0.8 => P(A′) = 1 – P(A) = 0.2
- P(B) = 0.6 => P(B′) = 1 – P(B) = 0.4
- P(C) = 0.3 => P(C′) = 1 – P(C) = 0.7

To find P(D).

The events corresponding to D are as follows:

- A∩B∩C′; A∩B′∩C; A′∩B∩C and A∩B∩C

$$P(D) = P(A\cap B\cap C') + P(A'\cap B\cap C) + P(A\cap B'\cap C) + P(A\cap B\cap C).$$

$$P(D) = P(A) \times P(B) \times P(C') + P(A') \times P(B) \times P(C) + P(A) \times P(B') \times P(C) + P(A) \times P(B) \times P(C)$$

$$= (0.8 \times 0.6 \times 0.7) + (0.8 \times 0.4 \times 0.3) + (0.2 \times 0.6 \times 0.3) + (0.8 \times 0.6 \times 0.3)$$

$$= 0.336 + 0.096 + 0.036 + 0.144 = 0.612$$

Hence, the probability that at least 2 of the 3 persons will remain alive at the end of 5 years is 0.612.

8.4 Bayes’ Theorem

The extension concept of conditional probability is Bayes theorem, which was introduced by Thomas Baye during 1700s. In this application of conditional probability, the stress is given on sequential events, especially when the information received from a second event is used to modify the probability that a first event has occurred.

Statement:

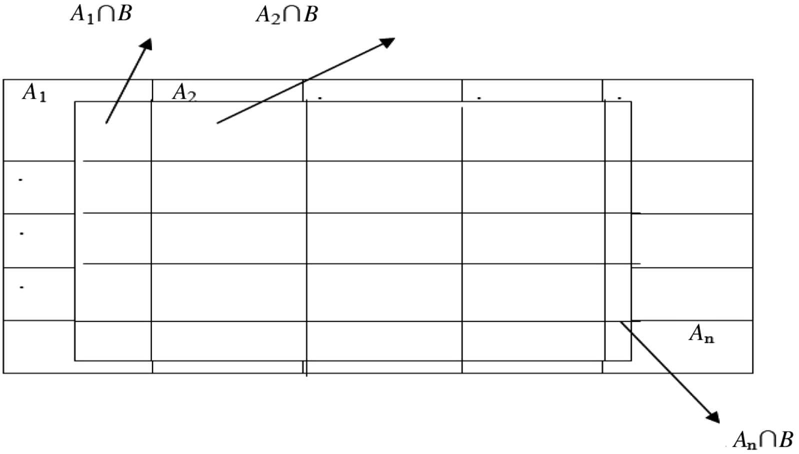
If $A_1, A_2, \ldots, A_n$ are mutually exclusive events with $P(A_i) > 0$;

(I = 1, 2, … , n), then for any event B which is a subset of $(A_1 \cup A_2 \cup \ldots \cup A_n)$

such that P(B) > 0, then

$$P(A_i/B) = \frac{P[A_i] \times P[B/A_i]}{\sum_{i=1}^{n}\left(P[A_i] \times P[B/A_i]\right)}; \quad i = 1,2, \ldots, n$$

The statement can be explained through a diagram.



Obviously, the events $A_1 \cap B$, $A_2 \cap B$, … , $A_n \cap B$ exist, and all are mutually exclusive.

$$\text{Then } B = (A_1 \cap B) \cup (A_2 \cap B) \cup \ldots \cup (A_n \cap B)$$

Then the probability of B can be given as $P(B) = P[(A_1 \cap B) \cup (A_2 \cap B) \cup \ldots \cup (A_n \cap B)]$

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \ldots + P(A_n \cap B)$$

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B) \tag{8.9}$$

By definition, $P(B/A_i) = P(A_i \cap B)/P(A_i)$; $P(A_i) > 0$, $i = 1, 2, \ldots, n$

Then, we have,

$$P(A_i \cap B) = P(A_i) \times P(B/A_i);\ i = 1, 2, \ldots, n \tag{8.10}$$

Using Equation (8.10) in Equation (8.9),

$$P(B) = \sum_{i=1}^{n} \left(P[A_i] \times P[B/A_i]\right) \tag{8.11}$$

By definition,

$$P(A_i/B) = P(A_i \cap B)/P(B);\ P(B) > 0;\ i = 1, 2, \ldots, n \tag{8.12}$$

Using Equations (8.10) and (8.11) in Equation (8.12),

$$P(A_i/B) = \frac{P[A_i] \times P[B/A_i]}{\sum_{i=1}^{n}\left(P[A_i] \times P[B/A_i]\right)};\ i = 1,2,\ldots, n$$

Hence, the theorem is proved.

Example:
In a bolt factory, machines X, Y, and Z manufacture 20%, 35%, and 45% of items, respectively. Out of which 8%, 6%, and 5% items are defective from machines Y and Z. One bolt is drawn at random from the product and is found defective. What is the probability that it is manufactured by machine Z?

Events

B_1: The bolt was manufactured by machine X.

B_2: The bolt was manufactured by machine Y.

B_3: The bolt was manufactured by machine Z.

Prior probability

This is an initial probability based on the prior level of information on the basis,

- $P(B_1) = 0.2$ because machine X produces 20% of the products.
- $P(B_2) = 0.35$ because machine Y produces 35% of the products.
- $P(B_3) = 0.45$ because machine Z produces 45% of the products.

Additional information

At the time of random selection, the selected bolt was found to be defective.

Event A: The selected bolt is defective.

Posterior probability

This is the revised probability that has the benefit of additional information. It is a conditional probability and can be expressed $P(A/B_i)$.

$P(A/B_1) = 0.08$; 8% bolt produced by machine X was defective.

$P(A/B_2) = 0.06$; 6% bolt produced by the machine X was defective.

$P(A/B_3) = 0.05$; 5% bolt produced by the machine X was defective.

Tabulate the prior and posterior probabilities:

Machine	Production (%)	$P(B_i)$ Prior Probability	Error (%)	$P(A/B_i)$ (Posterior Probability)	$P(B_i) \times P(A/B_i)$
X	20	0.2	8	0.08	0.0160
Y	35	0.35	6	0.06	0.0210
Z	45	0.45	5	0.05	0.0225
				Total	0.0595

To find P(B_3/A).

$$\text{By definition, } P(B_i/A) = \frac{P[B_i]\times P[A/B_i]}{\sum_{i=1}^{3}\left(P[B_i]\times P[A/B_i]\right)}; \quad i = 1,2,3$$

$$P(B_3/A) = \frac{P[B_3]\times P[A/B_3]}{\sum_{i=1}^{3}\left(P[B_i]\times P[A/B_i]\right)}; \quad i = 1,2,3$$

$$= 0.0225/0.0595 = 0.378.$$

The probability that the selected defective bolt manufactured by machine Z is 0.378 Representation in the form of tree diagram (Figure 8.1).

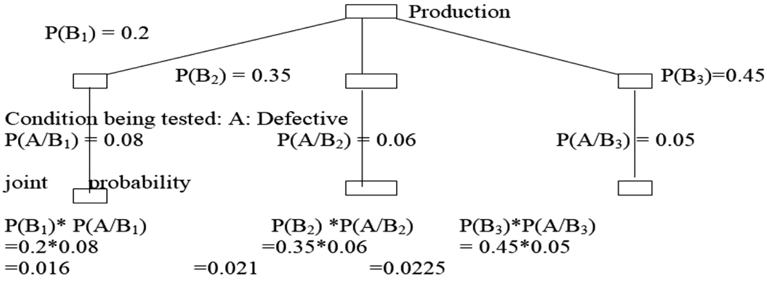


FIGURE 8.1
Tree Diagram.

Example:

A dryer manufacturer purchases heating elements from the different suppliers: Argostat, Bermrock, and Thermotek. Of those, Argostat provides 30%, Bermrock 50%, and Thermotek 20%. The elements are mixed in a supply bin before inspection and installation. Based on the past experience, 10% of the Argostat elements are defective, compared to only 5% of those supplied by Bermrock and just 4% of those by Thermtek. An assembly worker randomly selects an element for installation. What is the probability that the element was supplied by Argostat?

Events:

B_1: The heating element was supplied by Argostat.

B_2: The heating element was supplied by Bermrock.

B_3: The heating element was supplied by Thermtek.

A: The tested element is defective.

Tabulate the prior and posterior probabilities

Company	Supply (%)	$P(B_i)$	Error (%)	$P(A/B_i)$	$P(B_i) \times P(A/B_i)$
Argostat	30	0.3	10	0.10	0.030
Bermrock	50	0.5	5	0.05	0.025
Thermtek	20	0.2	4	0.04	0.008
				Total	0.063

$$\text{To find } P(B_1/A),\ P(B_i/A) = \frac{P[B_i] \times P[A/B_i]}{\sum_{i=1}^{3} \left(P[B_i] \times P[A/B_i]\right)}; \quad i = 1,2,3$$

$$P(B_1/A) = \frac{P[B_1] \times P[A/B_1]}{\sum_{i=1}^{3} \left(P[B_i] \times P[A/B_i]\right)}; \quad i = 1,2,3$$

$$= 0.030/0.063 = 0.476$$

The probability that the selected defective element was supplied by Argostat is 0.476

Example:

A person has 2 coins; one is unbalanced and lands heads 60% of the time, the other is fair and lands heads 50% of the time. He selects one of the coins and flips it. The result is heads.

1. What is the prior probability that the fair coin was selected?
2. Given additional information in the form of the single flip that came up as head, what is the revised probability that the coin is the fair one?

Event:

B_1: The selected coin was unbalanced.

B_2: The selected coin was fair.

A: To get head in a flip.

Tabulate the prior and posterior probabilities.

Coins	$P(A/B_i)$ Prior	$P(A/B_i)$ Posterior	$P(B_i) \times P(A/B_i)$
Unbalanced	0.5	0.6	0.30
Fair	0.5	0.5	0.25
		Total	0.55

1. $P(B_2) = 0.5$

$$P(B_i/A) = \frac{P[B_i]\times P[A/B_i]}{\sum_{i=1}^{2}\left(P[B_i]\times P[A/B_i]\right)}; \quad i = 1,2$$

2. $$P(B_2/A) = \frac{P[B_2]\times P[A/B_2]}{\sum_{i=1}^{2}\left(P[B_i]\times P[A/B_i]\right)}; \quad i = 1,2$$

$$= 0.25/0.55 = 0.455$$

Hence,

1. The prior probability of selection of a fair coin is 0.5.
2. The probability to get the head in a single flip using a fair coin is 0.455

Example:

There are 2 identical boxes containing 4 white and 3 red balls, 3 white and 7 red balls, respectively. A box is chosen at random, and a ball is drawn from it. If the ball is white, then what is the probability that it is from the first box?

Events:

B_1: Selection of the box 1;
B_2: Selection of the box 2;
A: Selection of white ball.

Box 1

4W	3R

Box 2

3W	7R

Total balls in box 1 = 7; Total balls in box 2 = 10

Selection of one ball from box 1 = $^7C_1 = 7$;

Number of white balls = 4; favourable chances of selection of 1 white ball = $^4C_1$ = 4

$$P(A/B_1) = 4/7 = 0.571$$

Selection of 1 ball from box 2 = $^{10}C_1 = 10$

Number of white balls = 3; favourable chances of selecting 1 white ball = $^3C_1$ = 3

$$P(A/B_2) = 3/10 = 0.3$$

Tabulate the prior and posterior probabilities

Box	$P(B_i)$	$P(A/B_i)$	$P(B_i) \times P(A/B_i)$
Box 1	0.5	0.571	0.286
Box 2	0.5	0.3	0.150
		Total	0.436

To find $P(B_1/A)$,

$$P(B_i/A) = \frac{P[B_i] \times P[A/B_i]}{\sum_{i=1}^{2} \left(P[B_i] \times P[A/B_i]\right)}; \quad i = 1,2$$

$$P(B_1/A) = \frac{P[B_1] \times P[A/B_1]}{\sum_{i=1}^{2} \left(P[B_i] \times P[A/B_i]\right)}; \quad i = 1,2$$

$$= 0.286/0.436 = 0.656$$

The probability of selection of a white ball from box 1 is 0.656.

Exercise 8

1. If 2 dice are thrown, what is the probability that the sum of numbers that appeared on them is (a) greater than 8? (b) neither 7 nor 11?
2. The probability that a student A solves a mathematics problem is 2/5, and the probability that a student B solves it is 2/3. What is the probability that the problem is not solved when they are working independently?
3. Based on a recent nationwide poll, a seller of printed advertisements estimates that 56% of all adults usually open all the mail they receive. If this is still the current rate at which adults open mail, what is the probability that, in a random sample of 1000 adults, the number who usually open all of their mail will be (a) less than 541? (b) 570 or more?
4. Among 1000 applicants for admission to an MBA program in a university, 600 were mathematics graduates and 400 were nonmathematics graduates. Of the mathematics graduate applicants, 30%, and of the no-mathematics graduates, 5%, were admitted. If an applicant selected at random is found to have been given admission, what is the probability that he or she is a mathematics graduate?
5. Consider a population of consumers consisting of 2 types. The upper class of customers comprises 35% of the population, and each member has probability 0.8 of purchasing brand A of a product. Each of the rest of the population has probability 0.3 of purchasing brand A. A consumer chosen at random is a buyer of brand A. What is the probability that the buyer belongs to the middle and upper class of consumers?

6. The members of a consulting firm rent cars from 3 rental agencies: 60% from Agency 1, 30% from Agency 2, and 10% from Agency 3. If 9% of the cars from agency 1 need a tune up, 20% of the cars from agency 2 need a tune-up, and 6% of the cars from agency 3 need a tune-up, what is the probability that a rental car delivered by the 2-firm will need a tune up? (Hint: Bayes application).
7. The manufacturer of a certain product has installed 3 machines, A, B, and C, and all are meant for producing a given product. All 3 machines are equally efficient and constitute 25%, 35%, and 40%, respectively, of a day's total production. It has been found that on an average, machine A produces 1% defective items, B produces 2% defective items, and C produces 3% defective items. An item is drawn at random from the combined output of all the 3 machines produced during a specified hour. Find the probabilities that the item selected is produced by (a) A, (b) B, and (c) C.
8. A manufacturing firm produces pipes in 2 plants, I and II, with daily production 1500 and 2000 pipes, respectively. The fraction of defective pipes produced by the two plants is 0.006 and 0.008, respectively. If a pipe selected at random from the day's production is found to be defective, what is the probability that it has come from plant I or plant II. (Hint: P[I] = 1500/3500 = 0.43; P[II] = 2000/3500 = 0.57)
9. A problem in business is given to 3 business students, S1, S2, and S3, whose chances of solving it are 0.6, 0.5, and 0.4, respectively. If they try it individually, what is the chance that the business problem will be solved?
10. Four balls are drawn at random from a bag containing 5 red and 7 blue balls. Compute the probability of getting (a) 4 red balls (b) 2 red and 2 blue balls, (c) 2 blue balls and 1 red ball.
11. Tech Search Inc. specializes in placing technical managers. It classifies clients in terms of skills and years of experience. The skills are research and development (R&D) and design. No one candidate possesses both skills. Experience categories are 2 years or less, between 2 and 10 years, and 10 years or more. At present, there are 100 executives on file with skills and experience summarized in the following table.

Experience	**R&D Skill**	**Design Skill**	**Total**
2 years or less	25	5	30
Between 2 and 10 years	15	15	30
10 years or more	5	35	40
Total	45	55	100

Suppose you select at random one executive's file. Determine each of the following probabilities:

a. P[R&D]

b. P[Design]

c. P[R&D and 10 years or more experience]

d. P[10 years or more experience R&D given an R&D executive is selected]

12. Five men in a company of 20 are graduates. If 3 men are picked out of the 20 at random, what is the probability that (a) They are all graduates? (b) There is no graduate? (c) What is the probability of at least 1 graduate?
13. A subcommittee of 6 members is to be formed out of a group consisting of 7 men and 4 ladies. Calculate the probability that subcommittee will consist of (a) exactly 2 ladies; (b) at least 2 ladies.
14. A construction company is bidding for 2 contracts, A and B. The probability that the company will get contract A is 3/5, the probability that the company will get contract B is 1/3, and the probability that the company will get both the contracts is 1/8. What is the probability that the company will get at least 1 contract?
15. The probability that a manager's job applicant has a postgraduate degree is 0.3 and has had some work experience as an office chief is 0.7 and that the applicant has both is 0.2. Out of 400 applicants, what number would have either a postgraduate degree or some professional work experience or both?
16. Mr. Sree Balaji is called for interview for 3 separate posts. At the first interview there are 5 candidates; at the second, 4 candidates; and at the third, 6 candidates. If selection of each candidate is equally likely, find the probability that Mr. Sree Balaji will be selected for (a) at least 1 post; (b) at least 2 posts.
17. A study by the Indian Energy Information Administration found that 84.3% of Indian households with income less than $10000 did not own a dishwasher, whereas only 21.8% of those with income greater than $50000 did not own a dishwasher. If one household is randomly selected from each group, determine the probability that
    a. neither household will own a dishwasher
    b. both households will have a dishwasher
    c. the lower-income household will own a dishwasher, but the higher-income household will not.
    d. the higher-income household will own a dishwasher, but the lower-income household will not.
18. A kitchen appliance has 16 working parts, each of which has a 0.99 probability of lasting through the product's warranty period. They operate independently, but with 1 or more malfunctions, the appliance will not work. What is the probability that a randomly selected appliance will work satisfactorily throughout the warranty period?
19. A taxi company in a small town has 2 cabs. Cab A stalls at red light 25% of the time, whereas Cab B stalls just 10% of the time. A driver randomly selects one of the cars for the first trip of the day. What is the probability that the engine will stall at the first red light the driver encounters?
20. A magician has 2 coins: one is unbalanced and lands heads 60% of the time, and the other is fair and lands heads 50% of the time. A member of the audience randomly selects one of the coins and flips it. The result is a head.
    a. What is the probability that the fair coin was selected?
    b. Given additional information in the form of the single flop that came up heads, what is the revised probability that the coin is a fair one?

21. Machine A produces 3% defective items, machine B produces 5% defective items, and machine C produces 10% defective items. Of the total output, 60% of the items is from machine A, 30% from B, and 10% from C, respectively. One of the items is selected randomly from a day's production.

    a. What is the prior probability that the item came from machine C?

    b. If the inspection finds the item to be defective, what is the revised probability that the item came from machine C?

9

Random Variables and Expectation

9.1 Introduction

We know that an experiment refers to an activity or measurement that results in an outcome. Clearly, tossing of a single coin is an experiment. When we toss a coin, we do not know whether it will turn heads or tail, and the chance of the head is half and the tail is half. An experiment is said to be a random experiment if its outcome depends on chance. A random variable can be defined based on random experiment. Usually the random variable is used to define the probability distribution and expectation.

9.2 Random Variable

A random variable is a variable that can take on different values according to the outcome of an experiment. It can be classified as follows:

- Discrete random variable or
- Continuous random variable

It is called 'random' because we do not know ahead of time exactly what value it will have following the experiment:

9.2.1 Discrete Random Variable

A random variable can take only certain values along an interval. In throwing dice, the outcome can be either 1 or 2 or 3 or 4 or 5 or 6.

x	1	2	3	4	5	6

Clearly the values of x are discrete.

9.2.2 Continuous Random Variable

A random variable can take any value in a given interval. Take the temperature measured of a location at a specific point of time. Clearly, the temperature can take any value. Usually the random variable will be denoted by X or Y.

9.3 Probability Distribution

A probability distribution is the relative frequency distribution that theoretically occurs for observations from a given population. Otherwise, it is a listing of all possible outcomes of an experiment, along with their respective probabilities of occurrence. It can be classified into

1. Discrete probability distribution or
2. Continuous probability distribution

9.3.1 Discrete Probability Distribution

If a random variable X assumes m different values say X_1, X_2, … , X_m with respective probabilities p_1, p_2, … , p_n ($p_i \geq 0$; $i = 1, 2, \ldots, n$, $p_1 + p_2 + \ldots + p_n = 1$), then the occurrence of the values X_i with their probabilities $p_i(i = 1, 2, \ldots, n)$ is called the 'discrete probability distribution'.

The same can be represented in the following tabular form:

X	X_1	X_2	…	X_i	X_{i+1}	…	X_n
P(x)	p_1	p_2	…	p_i	p_{i+1}	…	P_n

Example:

An experiment is conducted in which a fair coin is tossed (flipped) twice. The result of an experiment will be the random variable,

X = the number of times that heads comes up.

The sample space for this event can be defined as

$$S = \{HH, HT, TH, TT\}; n(S) = 4$$

Event:

A: Has exactly no head

B: Has exactly 1 head

C: Has exactly 2 heads

A = {TT}; n(A) = 1

B = {TH, HT}; n(B) = 2

C = {HH}; n(C) = 1

P(A) = n(A)/n(S) = ¼ = 0.25

P(B) = n(B)/n(S) = ½ = 0.5

P(C) = n(C)/n(S) = ¼ = 0.25

Then, the corresponding discrete probability distribution for the random variable, X = number of heads.

X	0	1	2
p	0.25	0.5	0.25

9.3.2 Characteristics of a Discrete Distribution

1. For any value of x, $0 \leq \mathrm{P}(x) \leq 1$.
2. The values of x are exhaustive. The probability distribution includes all possible values.
3. The values of x are mutually exclusive, only one value can occur for a given experiment.
4. The sum of their probabilities is 1. $\sum_{i=0}^{2} \mathrm{P}[x_i] = 1$

9.3.3 Probability Function

The probability function of the random variable X taking the value x can be defined as $f(x) = \mathrm{P}(X = x)$; where P refers to probability. It is otherwise called a 'probability mass function'.

$f(x)$ should satisfy the following two conditions:

1. $f(x) \geq 0$ for any value of x.
2. $\sum_{i=0}^{n} f[x_i] = 1$

In the previous example,

$$X = 0, f(0) = \mathrm{P}(X = 0) = 0.25$$

$$X = 1, f(1) = \mathrm{P}(X = 1) = 0.50$$

$$X = 2, f(2) = \mathrm{P}(X = 2) = 0.25$$

$f(0), f(1)$, and $f(2) \geq 0$

$$\sum_{i=0}^{2} f(i) = 1 = f(0) + f(1) + f(2) = 1$$

Example:

A financial counsellor conducts investment seminars with each seminar limited to 6 attendees. Because of the small size of the seminar group and the personal attention each person receives, some of the attendees became clients following seminar, for the past 20 seminars he conducted.

X = The number of visitors who became clients and has the following distribution:

X	0	1	2	3	4	5	6
P(x)	0.05	0.1	0.15	0.20	0.2	0.25	0.05

Find:

a. The probability that nobody will become a client.
b. The probability that at least 4 will become a client.

Given:

X	0	1	2	3	4	5	6
P(X)	0.05	0.1	0.15	0.2	0.2	0.25	0.05

a. $P(X = 0) = 0.05$

b. $P(X \geq 4) = P(X = 4) + P(X = 5) + P(X = 6) = 0.2 + 0.25 + 0.05 = 0.5$

Example:

X is a discrete random variable that has the following probability distribution.

X	0	1	2	3	4	5	6	7
P(X)	0	k	2k	2k	3k	k^2	$2k^2$	$7k^2+k$

Find (a) the value of k; (b) the value of $P(X > 6)$; and (c) the value of $P(X \geq 2)$

Given:

X	0	1	2	3	4	5	6	7
P(X)	0	k	2k	2k	3k	k^2	$2k^2$	$7k^2+k$

1. As a characteristic,

$$\sum_{x=0}^{7} P[x] = 1; \text{ Implies that, } 0 + k + 2k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k = 1$$

$$10k^2 + 9k = 1; 10k^2 + 9k - 1 = 0 \quad (9.1)$$

Equation (9.1) is a quadratic equation; it can have 2 values for k.

$$10k^2 + 10k - k - 1 = 0; 10k(k + 1) - 1(k + 1) = 0; (k + 1)(10k - 1) = 0$$

Hence, $k = -1, 1/10$

Because k is a component of probabilities, its value cannot be negative (i.e., $k = 1/10$). Then

X	0	1	2	3	4	5	6	7
P(X)	0	0.1	0.2	0.2	0.3	0.01	0.02	0.17

a. Find $P(X < 6)$

$$P(X < 6) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) = 0 + 0.1 + 0.2 + 0.2 + 0.3 + 0.01 = 0.81; P(X < 6) = 0.81$$

b. Find P($X \geq 2$)

$$\sum_{x=0}^{7} \mathrm{P}(x) = 1$$

$$\sum_{x=0}^{1} \mathrm{P}(x) + \sum_{x=2}^{7} \mathrm{P}(x) = 1$$

$$\sum_{x=2}^{7} \mathrm{P}(x) = 1 - \sum_{x=0}^{1} \mathrm{P}(x) = 1 - [\mathrm{P}(\mathrm{X} = 0) + \mathrm{P}(x = 1)] = 1 - [0 + 0.1] = 0.9$$

9.4 Mathematical Expectation

The mathematical expectation of the discrete probability is defined as,

X	x_1	x_2	…	x_i	x_{i+1}	…	x_n
P(X)	p_1	p_2	…	p_i	p_{i+1}	…	p_n

$$\mathrm{E(X)} = \sum_{x=0}^{n} x_i p_i = p_1\, x_1 + p_2\, x_2 + \dots + p_n x_n;\ \text{where } p_i \geq 0;\ \mathrm{i} = 1, 2, \dots, n \text{ and } \sum_{x=0}^{n} p_i = 1$$

Mean of a random variable: $\bar{\mathrm{A}} = \sum_{x=0}^{n} x_i p_i \,/ \sum_{x=0}^{n} p_i = 1 = \sum_{x=0}^{n} x_i p_i$

Hence, $\mathrm{E(X)} = \bar{x} = \text{Mean} = \sum_{x=0}^{n} x_i p_i$

Standard Results

$$\mathrm{E}(a) = a;\ \text{where } a \text{ is a constant}$$

$$\mathrm{E}(ax) = a \cdot \mathrm{E}(x);\ \text{where } a \text{ is a constant}$$

$$\mathrm{E}(x - \bar{x}) = 0$$

$$\mathrm{E}(x + y) = \mathrm{E}(x) + \mathrm{E}(y) \text{ where } x \text{ and } y \text{ are 2 discrete random variables.}$$

$$\mathrm{E}(X_1 + X_2 + \dots + X_n) = \sum_{x=0}^{n} \mathrm{E}[x_i];\ \text{where } X_1, X_2, \dots, X_n \text{ are } n \text{ discrete random variables.}$$

$$\mathrm{E}(x \times y) = \mathrm{E}(x) \times \mathrm{E}(y)$$

$$\mathrm{E}(a + bx) = a + b\mathrm{E}(x), \text{ where } a \text{ and } b \text{ are constants.}$$

Example:

Consider the following discrete probability distribution:

X	0	1	2	3	4	5	6
P(X)	0.05	0.1	0.2	0.25	0.15	0.15	0.10

Find E(X).

By definition, $E(X) = \sum_{x=0}^{6} x_i p_i = p_1 x_1 + p_2 x_2 + p_3 x_3 + p_4 x_4 + p_5 x_5 + p_6 x_6$

$$= 0 \times 0.05 + 1 \times 0.1 + 2 \times 0.2 + 3 \times 0.25 + 4 \times 0.15 + 5 \times 0.15 + 6 \times 0.1$$

$$= 0.05 + 0.1 + 0.4 + 0.75 + 0.6 + 0.75 + 0.6$$

$$= 3.2$$

Variance of a Random Variable
The variance of a random variable X can be defined as $E[(X - \bar{A})^2]$.
where $\bar{A} = E(X)$.

$$\text{Denoted by } Var(X) = E[(X - \bar{A})^2]$$

NOTE: $Var(X) = E[(X - \bar{A})^2] = E(X^2) - (E(X))^2$

Example:
A random variable x has the following probability distribution:

X	0	1	2	3
P(X)	1/7	3/7	2/7	1/7

Find the SD

Given:

X	**P(X)**	**X^2**	**X P(X)**	**X^2 P(X)**
0	1/7	0	0	0
1	3/7	1	3/7	3/7
2	2/7	4	4/7	8/7
3	1/7	9	3/7	9/7
		Total	10/7	20/7

$$Var(X) = E(X^2) - (E(X))^2$$

$$= \sum_{x=0}^{3} x_i^2 p_i - \left(\sum_{x=0}^{3} x_i p_i \right)^2$$

$$= 20/7 - (10/7)^2 = 20/7 - 100/49 = 140 - 100/49 = 40/49$$

$$\text{Var}(X) = 40/49 = 0.8163$$

$$\text{SD}(X) = \sqrt{0.8163} = 0.904$$

Example:

A discrete random variable can have the values $x = 3$, $x = 8$, and $x = 10$, and the respective probabilities are 0.2, 0.7, and 0.1. Determine the mean, variance and standard deviation.

Given:

X	P(*X*)	*X* P(*X*)	X^2	X^2 (*X*)
3	0.2	0.6	9	1.8
8	0.7	5.6	64	44.8
10	0.1	1	100	10.0
	Total	7.2		56.6

$$\text{Mean} = \bar{\text{A}} = \text{E}(X) = \Sigma\, \{X\, \text{P}(X)\} = 7.2$$

$$\text{Var}(X) = \text{E}(X^2) - (\text{E}(X))^2 = \Sigma\, X^2 \cdot \text{P}(X) - (7.2)^2 = 56.6 - (7.2)^2 = 56.6 - 51.84$$

$$\text{Var}(X) = 4.76;\ \text{SD}(X) = \sqrt{4.76} = 2.18174$$

Hence, the mean = 7.2; Var(X) = 4.76; and SD(X) = 2.182

Example:

A music shop is promoting a sale in which the purchases of a compact disc can roll a die, and then deduct a dollar from the retail price for each dot shown on the rolled die. It is equally likely that the die will come up any integer from 1 to 6. The owner of a music shop pays $5.00 for each compact disc and then prices $9.00. During this special promotion, what will be the shop's average profit per compact disc sold?

Given:

Purchasing price/disc = $5.00

Selling price/disc = $9.00

When the disc is rolled, the outcome is 1, 2, 3, 4, 5, and 6.
Any face can turn with the probability 1/6.

X	1	2	3	4	5	6
Discounted Price, $	8	7	6	5	4	3
P	1/6	1/6	1/6	1/6	1/6	1/6

$$\text{Mean selling price} = \sum_{x=0}^{6} x_i p_i = (8 + 7 + \ldots + 3) \times 1/6 = 33/6 = \$5.50$$

$$\text{Average profit/disc} = (\text{mean selling price}) - \$5.00 = \$5.5 - \$5.00 = \$0.5$$

During the special promotion, the shop gets $0.5 profit per disc.

Exercise 9

1. An investor is examining the possibility of investing in Alpha Mobile Company. Based on the past performance, he has broken the potential results of the investment into 5 possible customers with accompanying probabilities. The outcomes are annual rates of return on a single share of stock that currently costs $150/–. Find the expected value of return for investing a single share of Alpha Mobile.

Return on Investment ($)	0	10	15	25	50
Probability	0.2	0.25	0.3	0.15	0.1

   If the investor usually purchases stock whenever the expected rate of return exceeds 10%, will he purchase stock according to this data?

2. (a) Obtain the probability distribution of *X*. (b) Calculate the expected value of *X*. (c) if you pay $ 6 to get a card, find the probability that you will lose money and what is the actual loss?

Prize, *X*	$4000	1000	100	5	0
Number of Cards	1	3	95	425	4476

3. A random variable, *X*, has the following probability function:

Values of *X*	0	1	2	3	4	5	6	7
P(x)	0	k	2k	2k	3k	k^2	2k^2	7k^2 + k

   Then find (a) the value of k and (b) evaluate $P[X < 5]$; $P[X \geq 6]$ and $P[0 < x < 5]$

4. An unbiased coin is tossed 4 times. If y denotes the number of tails, from the distribution of x by writing down all possible outcomes calculate the expected value and variance of x.

Hint: x	0	1	2	3	4
P(x)	1/16	1/4	3/8	1/4	1/16

$$\text{Variance} = E[x^2]-[E[x]]^2$$

10

Discrete Probability Distribution: Binomial and Poisson Distributions

10.1 Introduction

We can define a probability distribution as the relative frequency distribution that should theoretically occur for observations from a given population. In business and other contexts, it can be helpful to proceed from

1. A basic understanding of how a natural process seems to operate in generating events, too.
2. Identifying the probability that a given event may occur.

By using a probability distribution as a model that represents the possible events and their respective likelihoods of occurrence, we can make more effective decisions and preparations in dealing with the events that the process is generating.

10.2 Binomial Distribution

Binomial distribution is one of the most widely used discrete distributions. It deals with consecutive trials, each of which has 2 possible outcomes. It relies on what is known as the 'Bernoulli Process'.

10.2.1 Characteristics of a Bernoulli Process

1. There are 2 or more consecutive trials.
2. In each trial, there are just 2 possible outcomes (success or failure).
3. The trials are independent.
4. The probability of success is constancy to all trials.

10.2.2 Definition of Binomial Distribution

The binomial distribution is defined as $P(X) = {}^{n}C_{x}\, p^{x}\, q^{n-x}$; $x = 0, 1, 2, \ldots, n$

Where, n is the number of trials, x is the number of successes, p is the probability of success, and q is the the probability of failure ($q = 1-p$).

The same can be expressed in a tabular form:

X	0	1	2	…	n
P(X)	q^n	${}^nC_1\, p^{n-1}q$	${}^nC_2\, p^{n-2}q^2$	…	p^n

From the table, it is clear that for $x = 1, 2, \dots, n$ gives the successive terms of the binomial expansion of $(p + q)^n = 1^n = 1$ $[p + q = 1]$. The 2 constants, p and n, are known as the parameters of the distribution.

NOTE: It is otherwise called a 'Bernoulli distribution' or 'finite discrete distribution'. (n is finite.)

10.2.3 Conditions of Binomial Distribution

1. Trials are independent and carried over under identical conditions for a fixed number of times.
2. There are only 2 possible outcomes, success and failure.
3. The success probabilities should be constant for all trials.

10.2.4 Properties of Binomial Distributions

1. It is a discrete probability distribution. The random variable X takes the values $0, 1, 2, \dots, n$, where n is finite.
2. Mean = np; variance = npq; standard deviation = $\sqrt{[npq]}$;

   $$\text{Skewness} = \frac{q-p}{\sqrt{npq}} \quad \text{and} \quad \text{Kurtosis} = \frac{1-6pq}{npq}$$
3. The mode corresponds to the value of x for which the P(X) is the maximum.
4. If $X(n_1, p)$ and $Y(n_2, p)$ are the 2 random variables following binomial distribution, then $(X + Y)$ with parameters $(n_1 + n_2, p)$ will be a random and follow binomial distribution.

10.2.5 Mean of Binomial Distribution

Show that the mean of binomial distribution is np.

$$\text{Mean} = E(X) = \sum_{x=0}^{n} x\text{P}[x]$$

$$= \sum_{x=0}^{n} \{x \times \{{}^nC_x\, p^x\, q^{n-x}\}\} = \sum_{x=0}^{n} \left\{ x \frac{n!}{x![n-x]!} p^x q^{n-x} \right\} = \sum_{x=0}^{n} \left\{ x \frac{n!}{x * [x-1]![n-x]!} p^x q^{n-x} \right\}$$

$$= \sum_{x=1}^{n} \left\{ \frac{n!}{[x-1]![n-x]!} p^x q^{n-x} \right\} = np \sum_{x=1}^{n} \left\{ \frac{[n-1]!}{[x-1]![n-x]!} p^{x-1} q^{n-1-[x-1]} \right\}$$

$$= np\{{}^{n-1}C_{x-1}\, p^{x-1}\, q^{n-1-[x-1]}\} = np\,(p+q)^{n-1} = np(1)^{n-1} = np.$$

10.2.6 Variance of Binomial Distribution

Show that the variance of binomial distribution is *npq*.

$$\text{By definition, variance}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2.$$

We know that E(*X*) = *np*. Consider E(*X*²).

$$\mathrm{E}(X^2) = \sum_{x=0}^{n} x^2 \mathrm{P}[x] = \sum_{x=0}^{n} x^2 \{{}^n\mathrm{C}_x p^x q^{n-x}\}$$

$$= \sum_{x=0}^{n}\left\{x^2 \frac{n!}{x![n-x]!} p^x q^{n-x}\right\} = \sum_{x=0}^{n}\left\{x^2 \frac{n!}{x[x-1]![n-x]!} p^x q^{n-x}\right\}$$

$$= \sum_{x=1}^{n}\left\{x \frac{n!}{[x-1]![n-x]!} p^x q^{n-x}\right\} = \sum_{x=1}^{n}\left\{[[x-1]+1] \frac{n!}{[x-1]![n-x]!} p^x q^{n-x}\right\}$$

$$= \sum_{x=1}^{n}\left\{[x-1] \frac{n!}{[x-1]![n-x]!} p^x q^{n-x}\right\} + \sum_{x=1}^{n}\left\{\frac{n!}{[x-1]![n-x]!} p^x q^{n-x}\right\}$$

$$= \sum_{x=1}^{n}\left\{[x-1] \frac{n!}{[x-1]*[x-2]![n-x]!} p^x q^{n-x}\right\} + np \quad = \sum_{x=1}^{n}\left\{\frac{n!}{[x-2]![n-x]!} p^x q^{n-x}\right\} + np$$

$$= \left\{n[n-1]p^2\right\} \sum_{x=1}^{n}\left\{\frac{[n-2]!}{[x-2]![[n-2]-[x-2]]!} p^{x-2} q^{[n-2]-[x-2]}\right\} + np$$

$$= (n(n-1)p^2)(p+q)^{n-2} + np = n(n-1)p^2 + np = n^2p^2 - np^2 + np = n^2p^2 + np(1-p)$$

$$E(X^2) = n^2p^2 + npq; \ \mathrm{Var}(X) = (n^2p^2 + npq) - (np)^2 = npq$$

$$\mathrm{Var}(x) = npq$$

NOTE: Standard derivation of binomial distribution, $\sigma = \sqrt{npq}$.

Example:

Researchers find that 60% of VCR owners understand how to program their VCR. Assuming a Bernoulli process and 3 randomly selected VCR owners, what is the probability of exactly 2 successes in 3 trials?

Given:

p = the probability that a VCR owner knows the VCR operations

$$n = 3; p = 0.6; q = 1-p = 0.4$$

To find P(X = 2). By definition P(X = x) = $\{^nC_x\, p^x\, q^{n-x}\}$

$$P(X = 2) = {}^3C_2\,(0.6)^2(0.4)^1 = 3 \times 0.36 \times 0.4 = 0.432$$

The probability of exactly 2 successes in 3 trials is 0.432.

Example:

Of the 41,636 residents of Tamil Nadu, 20% were born outside Tamil Nadu. A group of 5 people is to be randomly selected from the state and the discrete random variable is X, the number of persons in the group who were born in outside Tamil Nadu. Find

1. The probability for exactly 2 persons born outside Tamil Nadu.
2. The probability for at least 3 persons born outside Tamil Nadu.

Given:

p = 0.2 be the probability of the selected person to be born outside Tamil Nadu. Here, $n = 5$; $q = 1 - p = 1–0.2 = 0.8$.

1. $x = 2$

   By definition,

$$P(X = 2) = {}^5C_2\, p^2q^{5-2} = 10 \times (0.2)^2 \times (0.8)^3 = 0.2048 = 0.205$$

2. $x \geq 3$

   Clearly, the maximum value of x is 5.

$$P(x \geq 3) = P(X = 3) + P(X = 4) + P(X = 5)$$
$$= {}^5C_3(0.2)^3(0.8)^2 + {}^5C_4(0.2)^4(0.8) + {}^5C_5(0.2)^5$$
$$= 0.0512 + 0.0064 + 0.00032 = 0.05792 = 0.058$$

The probability for exactly 2 persons to be born outside Tamil Nadu is 0.205 and for 3 or more, 0.058.

Example:

The screws produced by a certain machine were checked by examining samples of 12. The following table shows the distribution of 128 samples according to the number of defective items they contained.

No. of Defects in a Sample of 12	0	1	2	3	4	5	6	7
No. of Samples	7	6	19	35	30	23	7	1

Use a binomial distribution, and find the expected frequencies if the chance of the machine being defective is ½. Find the mean and variance of the fitted distribution.

Let p be the probability of selection of defective screws, taken to be successful and $q = 1–p$.

Given: p = ½; q = 1–p = 1–1/2 = ½; n = 7 defective as maximum. Evaluate the number of samples with 0 defects, 1 defect, … , 7 defects out of 128 samples.

x	$P(x) = {}^7C_x(0.5)^x(0.5)^{7-x}$	Expected Frequency (128 × P(x))	fx	x^2	fx^2
0	1/128	1	0	0	0
1	7/128	7	7	1	7
2	21/128	21	42	4	84
3	35/128	35	105	9	315
4	35/128	35	140	16	560
5	21/128	21	105	25	525
6	7/128	7	42	36	252
7	1/128	1	7	49	49
			448	140	1792

$$\text{Mean} = \frac{\sum_{x=0}^{7} f_i x_i}{\sum_{x=0}^{7} f_i} = 448/128 = 3.5$$

$$\text{Variance} = \frac{\sum_{x=0}^{7} f_i x_i^2}{\sum_{x=0}^{7} f_i} - \left(\frac{\sum_{x=0}^{7} f_i x_i}{\sum_{x=0}^{7} f_i}\right)^2 = \left(1792/128\right) - (3.5)^2 = 14 - 12.25 = 1.75$$

$$\text{SD} = \sigma = \sqrt{1.75} = 1.323$$

Example:

If the probability of defective bolts is 1/10, find the following for the binomial distribution of defective bolts in a total of 400:

(a) mean, (b) variance, and (c) moment of skewness

Given:

n = 100; p = the probability for the selected bolt being defective = 1/10

NOTE: p refers the success probability; $q = 1 - p = 1 - 0.1 = 0.9$

By definition, mean = np = 400 × 0.1 = 40; variance = npq = 400 × 0.1 × 0.9 = 36

$$\text{MSK} = (q-p)/\sqrt{npq} = (0.9 - 0.1)/6 = 0.8/6 = 0.1333$$

10.3 Poisson Distribution

A Poisson distribution is a discrete probability distribution. This can be applied to events for which the probability of occurrence over a given span of time is extremely small. The discrete random variable, x, is the number of times the event occurs over the given span, and x can be 0, 1, 2, 3, … and so on, with theoretically no upper bound.

10.3.1 Definition of Poisson Distribution

The probability that an event will occur exactly x times over a given span of time is $\mathrm{P}[x]=\frac{e^{-\lambda}\times\lambda^{x}}{x!};x=0,1,2,\ldots,\infty$

where λ is the parameter and must be a positive constant, and $e = 2.71828$ (approximately).

Example:

A customer arrives at a service point during a given period of time, such as

1. The number of vehicles approaching the petrol bunk.
2. The number of persons entering a restaurant.
3. The number of calls received by a company switchboard.
4. The number of defects in manufacturing products.
5. The number of births, deaths, marriages, divorces, suicides, and other vital statistics over a given period of time.

10.3.2 Properties of Poisson Distribution

1. It is a discrete probability distribution in which the random variable, x, assumes the value $x = 0, 1, 2, \ldots, \infty$.
2. Mean = λ, variance = λ, $\sigma = \sqrt{\lambda}$, skewness = $1/\sqrt{\lambda}$, and kurtosis = $1/\lambda$; where λ is the parameter of the distribution.
3. The value of x corresponding to the maximum probability is taken to be the mode. It can have 1 or 2 modes. The number of modes can be decided based on the value of λ. If λ is an integer, then the 2 modes are $(\lambda - 1)$ and λ. If λ is not an integer, then the whole number lies between $(\lambda - 1)$, and λ is taken as mode.
4. If x and y are 2 independent Poisson variates with parameters λ_1 and λ_2, then the sum $(x + y)$ is also a Poisson variate with parameters $(\lambda_1 + \lambda_2)$.

10.3.3 Mean of the Poisson Distribution

Show that the mean of the Poisson distribution is λ. By definition,

$$\text{Mean}=\mathrm{E}(X)=\sum_{x=0}^{\infty}xP[x]=\sum_{x=0}^{\infty}\left\{x\times\left\{\frac{e^{-\lambda}\times\lambda^{x}}{x!}\right\}\right\}=\lambda\sum_{x=0}^{\infty}\left\{x\times\left\{\frac{e^{-\lambda}\times\lambda^{x-1}}{x\times[x-1]!}\right\}\right\}$$

$$=e^{-\lambda}\lambda\sum_{x=0}^{\infty}\left\{\frac{\lambda^{x-1}}{[x-1]!}\right\}=e^{-\lambda}\lambda\,e^{\lambda}=\lambda$$

Mean = λ

10.3.4 Variance of the Poisson Distribution

Show that the variance of the Poisson distribution is λ.

By definition,

$$\text{Variance}=\mathrm{E}(X^2)-(\mathrm{E}(X))^2=\mathrm{E}(X^2)-\lambda^2$$

$$\mathrm{E}(X^2)=\sum_{x=0}^{\infty}x^2\mathrm{P}[x]=\sum_{x=0}^{\infty}\left\{x^2\left\{\frac{e^{-\lambda}\times\lambda^x}{x!}\right\}\right\}=\lambda\sum_{x=1}^{\infty}\left\{x^2\left\{\frac{e^{-\lambda}\times\lambda^{x-1}}{x\times[x-1]!}\right\}\right\}$$

$$=e^{-\lambda}\lambda\sum_{x=1}^{\infty}\left\{x\left\{\frac{\lambda^{x-1}}{[x-1]!}\right\}\right\}=e^{-\lambda}\lambda\sum_{x=0}^{\infty}\left\{\left[[x-1]+1\right]\left\{\frac{\lambda^{x-1}}{[x-1]!}\right\}\right\}$$

$$=e^{-\lambda}\lambda\left\{\sum_{x=0}^{\infty}\left\{[x-1]\frac{\lambda^{x-1}}{[x-1]!}\right\}+\sum_{x=1}^{\infty}\left\{\frac{\lambda^{x-1}}{[x-1]!}\right\}\right\}$$

$$=e^{-\lambda}\lambda\left\{\sum_{x=2}^{\infty}\lambda\left\{\frac{\lambda^{x-2}}{[x-2]!}\right\}+\left\{e^{\lambda}\right\}\right\}=e^{-\lambda}\lambda\left\{\left[\lambda\, e^{\lambda}+e^{\lambda}\right]\right\}=\lambda^2+\lambda$$

$$\mathrm{E}\left(x^2\right)=\lambda^2+\lambda$$

$$\text{Variance}=[\lambda^2+\lambda]-\lambda^2=\lambda$$

Hence the variance and mean of the Poisson distribution are λ.

NOTE:

1. Mean = variance = λ, and SD = $\sqrt{\lambda}$
2. In a binomial distribution, if $n \to \infty$ and p becomes small, then it tends to be a Poisson distribution.
3. Whenever the value of λ is not given for a Poisson distribution, it can be approximately evaluated using the relation $\lambda = np$; where ($n \geq 20$) is the number of trial and $p \leq 0.05$ the probability of success.

Example:

For a discrete random variable that is Poisson distributed with $\lambda = 2$, evaluate the following: (1) P($X = 0$) (2) P($X \leq 2$) (3) P($X > 2$)

Given, $\lambda = 2$.

By definition,

$$\mathrm{P}[x]=\frac{e^{-\lambda}\times\lambda^x}{x!}=\frac{e^{-2}\times 2^x}{x!}$$

1. When $x = 0$; P($X = 0$) = $\dfrac{e^{-2}\times 2^0}{0!} = 0.135$ ($2^0 = 1$ and $0! = 1$)
2. When $X \leq 2$, P($X \leq 2$) = P($X = 0$) + P($X = 1$) + P($X = 2$)

$$=\frac{e^{-2}\times 2^0}{0!}+\frac{e^{-2}\times 2^1}{1!}+\frac{e^{-2}\times 2^2}{2!}=0.135+4\,e^{-2}=0.135+0.135\times 4=0.675$$

3. When $X > 2$, P($X > 2$) = P($X = 3$) + P($X = 4$) + … + P($X = \infty$)

We know that,

$$\sum_{x=0}^{\infty} \mathrm{P}[X = x] = 1$$

$$\sum_{x=0}^{\infty} \mathrm{P}[X = x] = \sum_{x=0}^{2} \mathrm{P}[X = x] + \sum_{x=3}^{\infty} \mathrm{P}[X = x] = 1$$

$$\sum_{x=3}^{\infty} \mathrm{P}[X = x] = 1 - \sum_{x=0}^{2} \mathrm{P}[X = x] = 1 - 0.675 = 0.325$$

Hence, $\mathrm{P}(X = 0) = 0.135$; $\mathrm{P}(X \leq 2) = 0.675$, and $\mathrm{P}(X > 2) = 0.325$.

Example:

In the year 2004, there were about 530 motor vehicle thefts for every 100,000 registrations. Assuming that

1. A Poisson distribution,
2. A community with a comparable theft rate and 1000 registered motor vehicles, and
3. x = number of vehicles stolen during the year in that community.

Determine the following: (1) E(X) (2) $p(X = 3)$ (3) $p(3 \leq X \leq 5)$

Let p be the probability of the theft.

Given $p = 530/100000 = 0.00530 = 0.0053$; $n = 1000$.

1. By definition,

$$\mathrm{E}(X) = \text{mean} = np = 1000 \times 0.0053 = 5.3 \text{ or } \lambda = 5.3$$

2. By definition,

$$\mathrm{P}(X = 3) = \mathrm{P}[X] = \frac{e^{-\lambda} \times \lambda^{x}}{x!} = \frac{e^{-5.3} \times 5.3^{3}}{3!} = 0.124$$

3. $\mathrm{P}(3 \leq X \leq 5) = \mathrm{P}(X = 3) + \mathrm{P}(X = 4) + \mathrm{P}(X = 5)$

$$= 0.124 + \frac{e^{-5.3} \times 5.3^{4}}{4!} + \frac{e^{-5.3} \times 5.3^{5}}{5!} = 0.124 + 0.164 + 0.174 = 0.462$$

Example:

Over the past year, a university's computer system has been struck by a virus at an average rate of 0.4 viruses per week. The university's information technology manager estimates that each time a virus occurs, it costs the university $1000 to remove the virus and repair the damages it has caused. Assuming a Poisson distribution, what is the probability that the university will have the good fortune of being virus free during the upcoming week? During this same week, what is the expected amount of money that the university will have to spend for virus removal and repair?

Given: Mean rate of repair = 0.4 (i.e., $\lambda = 0.4$)

Cost to rectify each attack = \$1000

To find (a)

$$P(X = 0) = \frac{e^{-0.4} \times 0.4^0}{0!} = 0.67$$

(b) The expected number of virus attack in that week = E(X) = λ = 0.4

Expected cost of repair in that week = 0.4 × 1000 = \$400

Hence, the probability of being virus attack free is 0.67

Expected cost of rectification is \$400.

Example:

Out of the total bulbs manufactured by a company, 5% bulbs are found to be defective. Use a Poisson distribution to find the probability that in a sample of 100 bulbs (1) none is defective and (2) 5 bulbs will be defective.

Let p be the event of manufactured bulbs being defective. Given

$p = 0.05 \Rightarrow q = 1 - p = 0.95$; $n = 100$; $\lambda = np = 5$.

To find (1) P(X = 0) (2) P(X = 5)

By definition,

$$P[x] = \frac{e^{-\lambda} \times \lambda^x}{x!}$$

1. $p(X = 0) = \dfrac{e^{-5} \times 5^0}{0!} = 0.0067$
2. $p(X = 5) = \dfrac{e^{-5} \times 5^5}{5!} = 0.1755$

The probability of none is defective is 0.0067 and that exactly 5 are defective is 0.1755.

Example:

If a random variable, X, follows a Poisson distribution such that $p(X = 1) = p(X = 2)$, find the mean and variance. Find also $p(X = 0)$.

By definition,

$$P[x] = \frac{e^{-\lambda} \times \lambda^x}{x!}$$

Given that P(X = 1) = P(X = 2), to find (1) mean, (2) variance, and (3) P(X = 0)

Since P(X = 1) = P(X = 2), $\dfrac{e^{-\lambda} \times \lambda^1}{1!} = \dfrac{e^{-\lambda} \times \lambda^2}{2!}$; $[\lambda/1] = [\lambda^2/2]$

Implies that $\lambda = 2$.

1. Mean = $\lambda = 2$

2. Variance = $\lambda = 2$

3. $P(X = 0) = \dfrac{e^{-2} \times 2^0}{0!} = 0.1353.$

Exercise 10

1. An insurance salesman sells policies to 5 men of identical age and good health. According to actuarial tables, the probability that a man of this particular age will be alive for 30 years hence is 2/3. Find the probability that for 30 years hence (a) at least 1 man will be alive, and (b) at least 3 men will be alive.
2. A car-hire firm has 2 cars that it hires out daily. The number of demands for a car on each day is distributed as a Poisson variate with mean 1.5. Calculate the proportion of days in which (a) neither car is used and (b) some demand is refused.
3. The number of road accidents on a highway during a month follows a Poisson distribution with mean 6. Find the probability that in a certain month number of accidents will be: (a) not more than 3 and (b) between 2 and 4. [Given: $e^{-6} = 0.00248$.]
4. Of the blades produced by a blade manufacturing factory, 1/5 turns out to be defective. The blades are supplied in packets of 10. Use a Poisson distribution to calculate the approximate number of packets containing no defective, 1 defective and 2 defective blades, and 1 defective blade, respectively, in a consignment of 100,000 packets. [Given: $e^{-0.02} = 0.9802$.]
5. Assuming that 4% of the output of a factory making certain parts is defective and that 100 units are in a package, what is the probability that at the most 3 defective parts may be found in a package?
6. The number of road accidents on a highway during a month follows a Poisson distribution with mean 6. Find the probability that in a certain month number accidents will be: (a) not more than 3 and (b) between 2 and 4.
7. The Bhavana Shree company, which manufactures medicine bottles, finds that 1% of the bottles are defective. The bottles are packed in boxes containing 500 bottles. A drug manufacturer buys 100 boxes from the producer of bottles. Using a Poisson distribution, find how many boxes will contain: (a) no defectives and (b) at least 2 defective bottles.

11

Continuous Probability Distribution: Normal Distribution

11.1 Introduction

This chapter deals with the probability distributions for continuous random variables, which can take any value in a given interval. This can be expressed as smooth curves, where the probabilities are expressed as areas under the curve.

11.2 Definition of Normal Distribution

A normal distribution is the most important continuous distribution in statistics. It is so important because

1. Many natural and economic phenomena tend to be approximately normal.
2. It can be used as a tool to approximate other distributions, which includes binomial and Poisson.
3. The sample means and proportions tend to be normally distributed.

It is defined by the probability density function,

$$f[x] = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\left(\frac{1}{2}\right)\left(\frac{x-\mu}{\sigma}\right)^2}; 0 < x < \infty$$

Here μ and σ stand for the mean and standard deviation. The curve representing this is referred, to as normal curve. [Refer Figure 11.1]

The total area bounded by the curve and the *x*-axis is

$$\int_0^\infty f[x]dx = 1$$

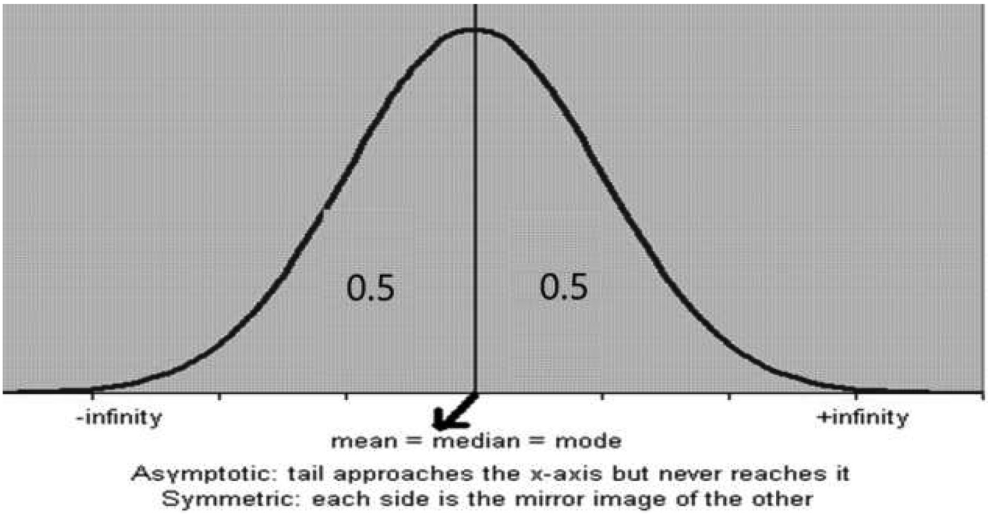


FIGURE 11.1
Outlook of the Normal Curve.

The area under the normal curve between the ordinates $x = c$ and $x = d$, where $c < d$, implies that probability that x lies between c and d, that is, $P(c < x < d)$.

The curve is symmetrical about the mean line. That is, each side is the mirror image of the other.

Refer Figure 11.2 for area under the curve.

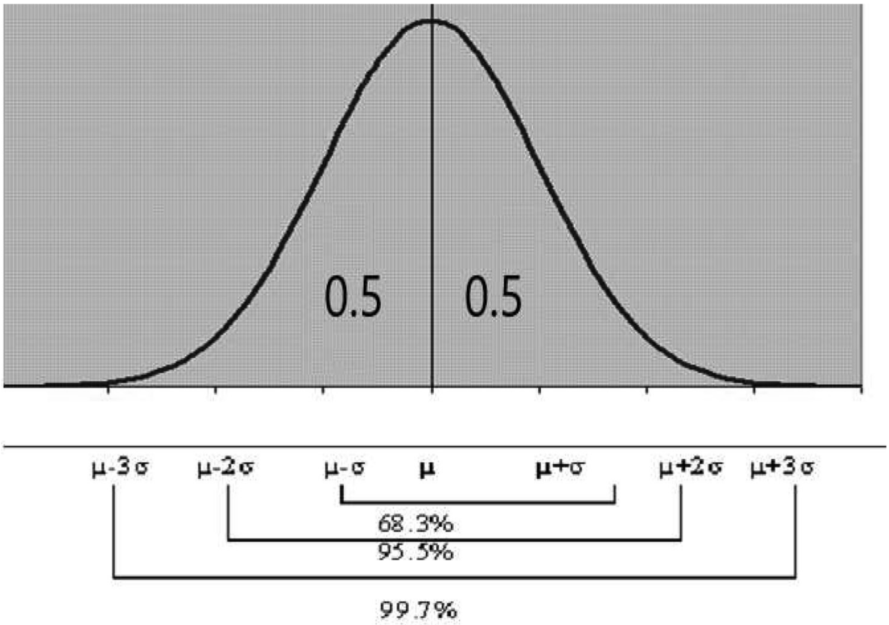


FIGURE 11.2
Area under the Normal Curve.

11.3 Standard Normal Distribution

If we take $Z = [x - \mu]/\sigma$, then $f[x] = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\left(\frac{1}{2}\right)(Z)^2}$;$0 < x < \infty$

$$\text{Mean} = \mathrm{E}(Z) = \mathrm{E}([x - \mu]/\sigma) = [1/\sigma]\ \{\mathrm{E}[X] - \mu\} = [1/\sigma]\ \{\mu - \mu\} = 0$$

$$\text{Variance of } Z = \mathrm{E}(Z^2) - (\mathrm{E}(Z))^2$$

$$= \mathrm{E}[([x - \mu]/\sigma)^2] = [1/\sigma^2][\mathrm{E}(x^2) + \mathrm{E}(\mu^2) - \mathrm{E}(2x\mu)] = 1/\sigma^2\ [\mathrm{E}(x^2) - \mu^2 - 2\mu^2]$$

$$= 1/\sigma^2\ [\mathrm{E}(x^2) - \mu^2] = [1/\sigma^2]\ [\mathrm{E}(x^2) - (\mathrm{E}(x))^2] = 1/\sigma^2 \times \sigma^2 = 1$$

Z is the standard normal variate with mean 0 and variance 1. It is denoted by $N\,\tilde{}\,(0,1)$.

Refer to Figure 11.3 for area under the standard normal curve.

Let $\varphi[z]$ stand for the area under the normal curve to the left of the ordinate Z (Figure 11.4).

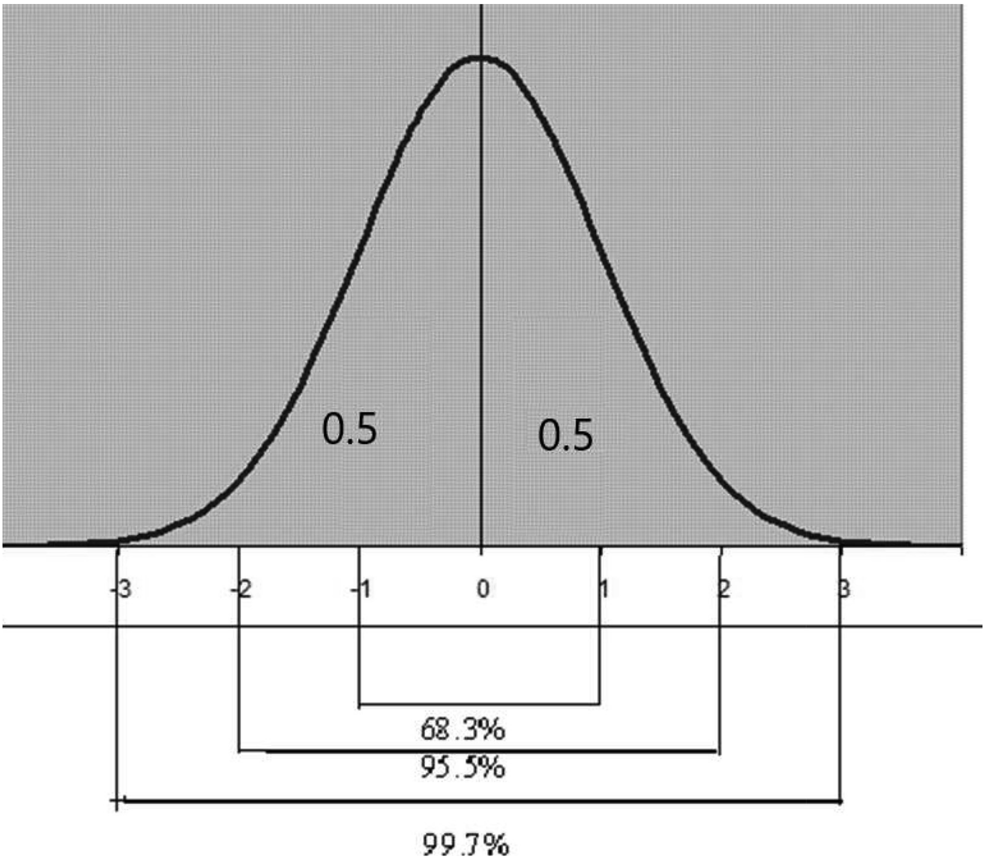


FIGURE 11.3
Area under the Standard Normal Curve.

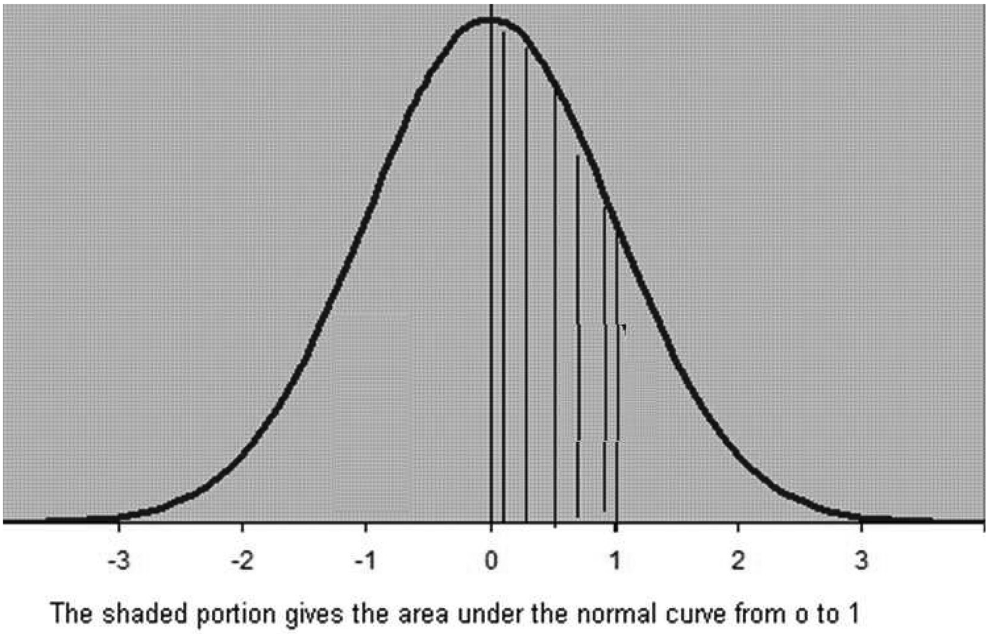


FIGURE 11.4
Area Specified for the Standard Normal Curve from $z = 0$ to 1.

$$\varphi[z] = \int_a^b \mathrm{P}[z]dz$$

$$\text{Because P}(z)\text{ is symmetrical, } \int_0^b \mathrm{P}[z]dz = \int_{-b}^0 \mathrm{P}[z]dz$$

From the standard table, the area from $z = 0$ to $z = 1$ is 0.3413.

11.4 Properties of Normal Distribution

1. It is a continuous probability distribution with μ and σ being the 2 parameters
2. Mean = μ; variance = σ^2; skewness = 0; kurtosis = 0; and
3. mean = median = mode = μ.
4. The curve is symmetrical about the mean line $x = \mu$ and passes through the peak of the curve. It separates the area into 2 equal parts.

Example:

Media researchers report the average daily TV viewing time for adult males to be 4.28 hours. Assume a normal distribution with a standard deviation of 1.30 hour;

what is the probability that a randomly selected adult male watches TV less than 2 hours per day?

Given: Mean = $\mu = 4.28$ hours; SD = $\sigma = 1.30$ hours.

Find P($x < 2$ hours)

Given: $x = 2$, we know that $Z = \{x - \mu\}/\sigma = (2 - 4.28)/1.30 = -1.75$

$$P(x < 2) = P(Z < -1.75);\ P(x < -1.75) = 0.5 - \varphi(-1.75) = 0.5 - 0.4599 = 0.0401$$

[Refer Figure 11.5.]

The probability of randomly selected adult male watches TV for less than 2 hours is 0.0401.

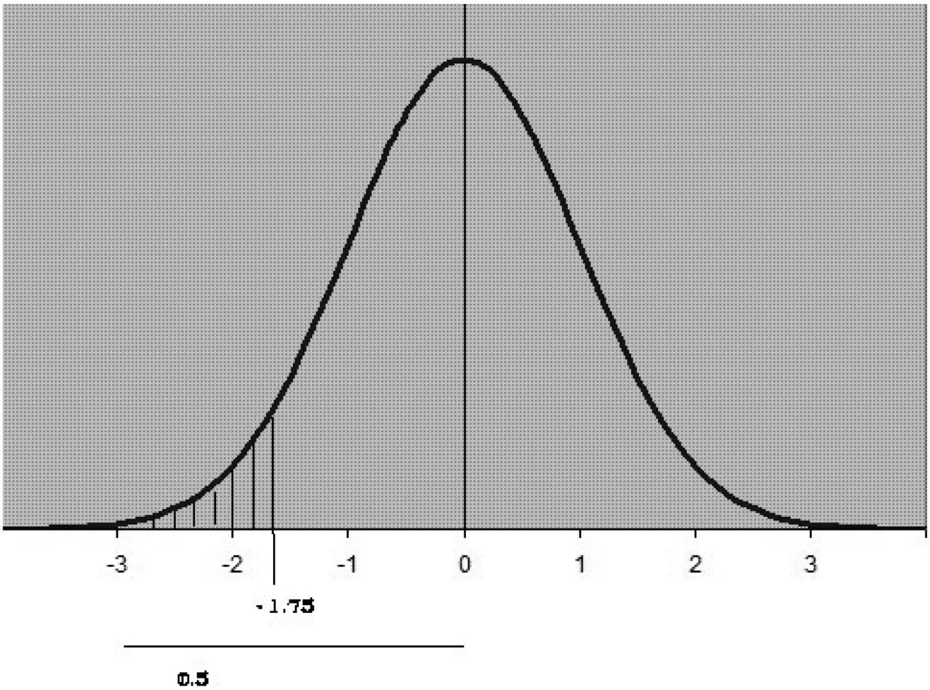


FIGURE 11.5
Area Specified for the Standard Normal Curve from $z = -3$ to -1.75.

Example:

Following their production, industrial generator shafts are tested for static and dynamic balance and the necessary weight is added and predrilled holes in order to bring each shaft within balance specifications. From the past experience, the amount of weight added to a shaft has been normally distributed with an average of 35 grams and a standard deviation of 9 grams. What is the probability that a randomly selected shaft will require between 35 and 40 grams of weight for proper balance? What is the probability that a randomly selected shaft will require at least 50 grams of weight for proper balance?

Given: Mean = $\mu = 35$ grams; SD = $\sigma = 9$ grams.

Find (1) $P(35 \le x \le 40)$ (2) $P(x \ge 50)$.

We know that $Z = (x - \mu)/\sigma$

1. When $x = 35$; $Z = (35 - 35)/9 = 0$.

$$\text{When } x = 40;\ Z = (40 - 35)/9 = 5/9 = 0.56$$

$$P(35 \le x \le 40) = P(0 \le Z \le 0.56) = 0.2123 \text{ (using tables)}$$

2. When $x = 50$; $Z = (50 - 35)/9 = 15/9 = 1.67$

$$P(x \ge 50) = P(Z \ge 1.67) = 0.5 - \varphi(Z = 1.67) = 0.5 - 0.4525 = 0.0475$$

[Refer Figure 11.6]

Hence,

1. The probability that the randomly selected shaft will require between 35 and 40 grams is 0.2123.
2. The probability that the randomly selected shaft will require at least 50 grams is 0.0475.

Example:

A sample of 100 dry battery cells tested to find the length of life produced the following results: $\mu = 12$ hours $\sigma = 3$ hours

Assuming the data are normally distributed, what percentage of battery cells is expected to have a life:

(1) >15 hours, (2) <6 hours, and (3) 10 and <14 hours

Given: Mean = $\mu = 12$ hours; SD = $\sigma = 3$ hours

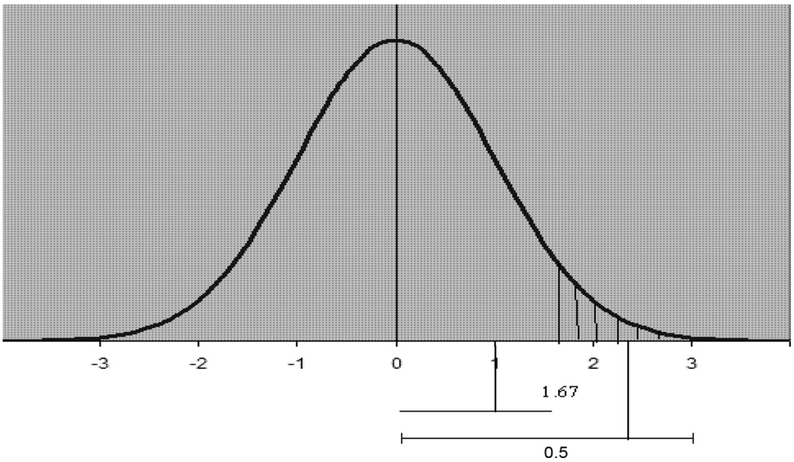


FIGURE 11.6
Area Specified for the Standard Normal Curve from $z = 1.67$ to 3.

(1) P(X > 15 hours), (2) P(X < 6 hours), (3) P(10 < X < 14 hours)

1. When $X = 15$; $Z = (x - \mu)/\sigma = (15 - 12)/3 = 1.0$

$$P(X > 15 \text{ hours}) = P(Z > 1) = 0.5 - \varphi(0 \leq Z \leq= 1) = 0.5 - 0.3413 = 0.1587$$

[Refer Figure 11.7]

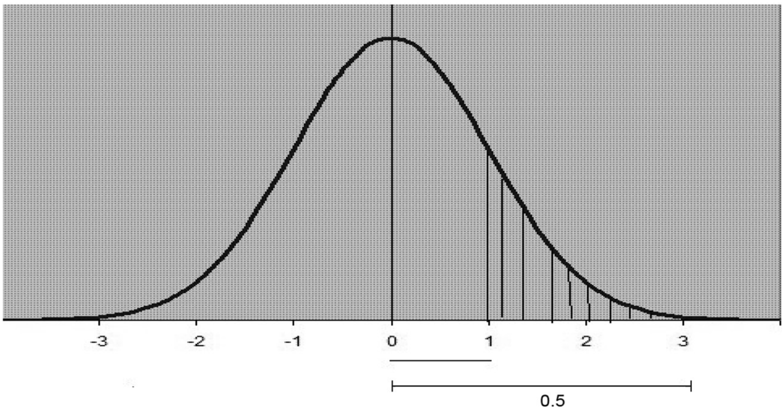


FIGURE 11.7
Area Specified for the Standard Normal Curve from $z = 1$ to 3.

2. When $X = 6$.

$$Z = (6 - 12)/3 = -2; P(X < 6) = P(Z < -2);$$

$$P(Z < -2) = 0.5 - \varphi\ (0 \leq Z \leq -2) = 0.5 - 0.4772 = 0.0228$$

Refer Figure 11.8.

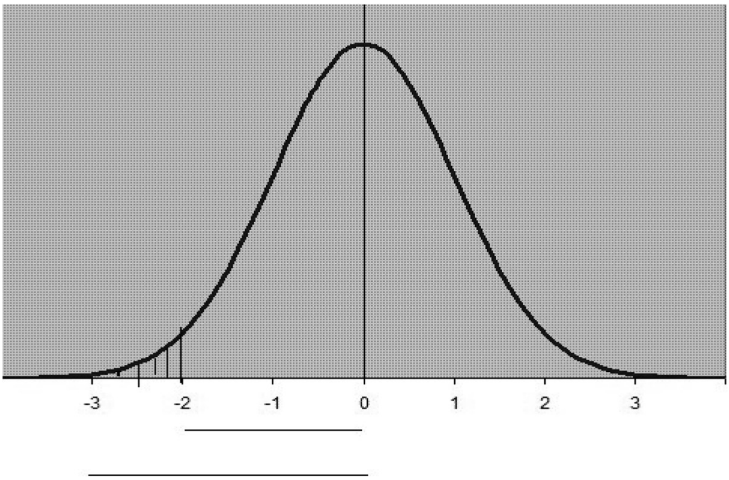


FIGURE 11.8
Area Specified for the Standard Normal Curve from $z = -3$ to -2.

3. When $X = 10$; $Z = (10 - 12)/3 = -2/3 = -0.67$; When $X = 14$; $Z = (14 - 12)/3 = 2/3 = 0.67$

$$P(10 < X < 14) = P(-0.67 < Z < 0.67) = = 2 \times P(0 < Z < 0.67) = 2 \times 0.2486 = = 0.4972$$

Refer Figure 11.9 because it is symmetric.

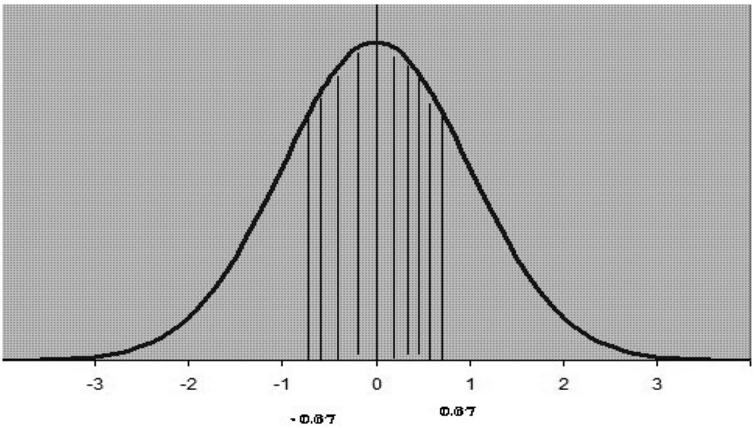


FIGURE 11.9
Area Specified for the Standard Normal Curve from $z = -1$ to $+1$.

Hence the required probabilities are

1. $P(X > 15 \text{ hours}) = 0.1587$
2. $P(X < 6 \text{ hours}) = 0.0228$
3. $P(10 < X < 14 \text{ hours}) = 0.4972$

Example:

In a college, the average score on the mathematics portion was 511; 21.77% of the students secured more than 600. Find σ.

Given: $\mu = 511$; $P(X \geq 600) = 0.2177$.

Find σ.

We know that $Z = (x - \mu)/\sigma = (600 - 511)/\sigma = 89/\sigma$ (11.1)

$$P(X \geq 600) = P(Z \geq 89/\sigma) = 0.2177 = 0.5 - 0.2177 = 0.2823$$

From the table, $0.2823 = P(0 \leq Z \leq 0.78) \Rightarrow Z = 0.78$ (11.2)

Using (11.2) in (11.1), $0.78 = 89/\sigma$; $\sigma = 89/0.78$; $\sigma = 114.10$

Hence, the required value of the SD is 114.10

Exercise 11

1. Ms. Bhavana Shree, who is good at credit scoring, has just completed her MBA and been appointed area manager of the bank of Maha. The newly appointed area manager is interested in issuing different varieties of credit cards with special schemes to families with different income levels. It was found in her study that there are 10,000 families in her area that fall under the new scheme, out of which a sample of 100 managers are taken for study.

   The frequency distribution of the survey of various income levels per month and corresponding number of families is as follows:

Income Level Per Month ($ in thousands)	10–15	15–20	20–25	25–30	30–35	35–40
Number of Families	11	20	35	20	8	6

   Consider a normal distribution in finding the following:

   a. Number of families whose salary is more than 30000/–

   b. Percentage of families falling under 15000/– to 25000.

   c. Number of families whose income levels are less than 10000/–

2. The number of calories in a salad on the lunch menu is normally distributed with mean 200 and SD 5. Find the probability that the salad you select will contain: (a) more than 208 calories and (b) between 190 and 200 calories.

3. The average number of units produced by a manufacturing concern per day is 355 with a standard deviation of 50. It makes a profit of $1.50 per unit. Determine the percentage of days when its total profit per item is: (a) between $457.50 and $645.00, (b) greater than $682.50, (c) between 305 and 430 units, and (d) greater than 455 units

4. The customer accounts of a certain department store have an average balance of $120 and standard deviation of $40. Assuming that the account balances are normally distributed:

   a. What proportion of accounts is more than $150?

   b. What proportion of accounts is between $100 and $150?

   c. What proportion of accounts is between $60 and $0?

5. A sales tax officer has reported the average sales of the 500 firms that he has to deal with during a year amount to $72,000 with a standard deviation of $20,000. Assuming that the sales in these firms are normally distributed, find

   a. The number of firms whose sales are more than $80,000 and

   b. The percentage of firms whose sales are likely to range between $60,000 and $80,000.

6. The diameters of ball bearings are normally distributed with mean 0.6140 inches and SD 0.0025 inches. Calculate the percentage of ball bearings with diameters: (a) between 0.610 and 0.618 inches, (b) greater than 0.617 inches, and (c) less than 0.608 inches.

7. A certain hospital usually admits 50 patients per day. On an average, 4 patients in 100 require facilities found in special rooms of the hospital. On the morning of a certain day, it is found that there are 4 such rooms available. Assuming that 50 patients will be admitted, find the probability that more than 4 patients will require such special rooms. (Note: here $n = 50$; $p = 4/100 = 0.04$, $m = np = 2$)

12

Theory of Sampling

12.1 Introduction

In this chapter, we discuss the concepts of sampling and sampling distributions, which are the actual basis of statistical estimation and hypothesis testing. The main purpose of sampling is to allow us to make use of the information gathered from the sample to draw influences about the entire population. One can define a 'population' as a collection of objects having a certain well-defined set of attributes. A 'sample' is any subset of a given population. It is possible to estimate the population parameters from the limited sample parameters with the help of statistical methods and concepts. This falls under the category of statistical inference (i.e., inductive statistics). The inferential process is not error free because the estimation or inference is based on the limited sample data obtained from samples.

We should evaluate such errors to have a measure of confidence in our inferences. If we take random samples, these errors occur randomly, and thus, the same can be computed probabilistically.

In this chapter, the concepts of sampling will be developed, sampling distributions for various sample statistics like the sample mean and proportion are described, and the well-known sampling distributions as the Chi-square, F-distribution, t-distribution, and standard normal distribution are also introduced. These distributions fit well into certain sample statistics that play a major role in estimation and hypothesis testing.

12.2 Why Sample?

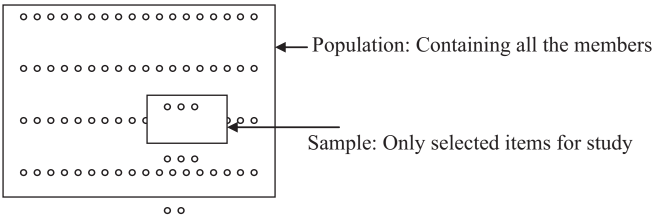
In many situations, even though we are interested in some characteristic of a specific population, we cannot physically examine the entire population because of cost, time, or other limitations. In such instances, we examine a part of a population by means of a sample with the expectation that it will be the representative of the population under study.

12.3 How to Choose It?

One way is to use simple random sampling. Simple random sampling provides all the samples of the size specified with an equal chance of being selected. Based on the given random sample, one can find a sample statistic such as the mean or variance and the same can be used to estimate the corresponding population parameter. Every statistic is a random variable with its own probability distribution. The probability distribution referred to by the sample statistic is known as a 'sampling distribution'. It has a defined property like any probability model. Based on the properties, one can evaluate the chance errors involved in drawing the inference from a sample.

12.4 Sample Design

A sample design is a procedure or plan for obtaining a sample from a prescribed population prior to collecting any data.


Population: Containing all the members
Sample: Only selected items for study

12.5 Keywords and Notations

Population: A collection of objects having certain well-defined set of attributes.

Example:

- The population of affiliated colleges in Tamil Nadu.
- The population of government hospitals in Tamil Nadu.

Sample: A portion of the population.

Example:

- Collection of affiliated colleges in Tamil Nadu with minority status.
- Collection of government hospitals only in Chennai.

Parameter: The characteristics of the population.

Example:

- The population mean, population SD, etc.

Statistic: The characteristics of the sample.

Example: sample mean, sample SD, etc.

Degrees of freedom: It means the number of items to be selected freely out of *n* items. It is ($n - 1$). It is denoted by *df.*

Example: Select 3 integer numbers in such a way that their addition leads to the value 100.

$$40 + 10 + 50 = 100$$

one can choose freely 2 items only, the selection of third value cannot be done freely. If you select 40 and 10; the third value should be 50.

Degrees of freedom = *df* = 3 − 1 = 2.

Census: The complete enumeration of the population.

Notations:

N	population size
μ	population mean
σ	population SD
p	population proportion
n	sample size
$\bar{x}$	sample mean
s	sample SD
p	sample proportion
R	population correlation coefficient
r	sample correlation coefficient

Sample survey: Process of partial enumeration is called a 'sample survey'.

12.6 Advantages and Disadvantages of Sampling

Advantages

1. Less time is needed to study the sample than the population.
2. With less cost toward the analysis in most numbers of situations, sampling gives adequate information.
3. The confidence level of data collected is more in sampling than in population.

Disadvantages

1. At times, there is a possibility of the error factor.
2. High degree of expertise is required while selecting the sample.

12.7 Nonrandom Errors and Non-sampling Errors

This type of error can occur in 2 different situations:

1. The sample is not selected from the corresponding population.
2. A sample is taken from a predefined population, but response bias, that is, respondents are not giving the proper information, is a factor.

12.8 Random Errors and Sampling Errors

At times, a well-designed sample may not provide an actual representative of the population under study; it is because a sample is a portion of a population. The inference based on this sample toward the parent population leads to incorrect inferences.

Such type of errors are known as ‘random errors’ or ‘sampling errors’.

12.9 Types of Samples

A sample can be classified into 2 major categories:

- Probability sample
- Nonprobability sample

12.9.1 Probability Sample

If the probability of selection of each member into a sample is non-zero, then the resulting sample is said to be a probability sample.

12.9.2 Nonprobability Sample

If a sample is not probabilistic sample, then it is said to be nonprobabilistic sample.

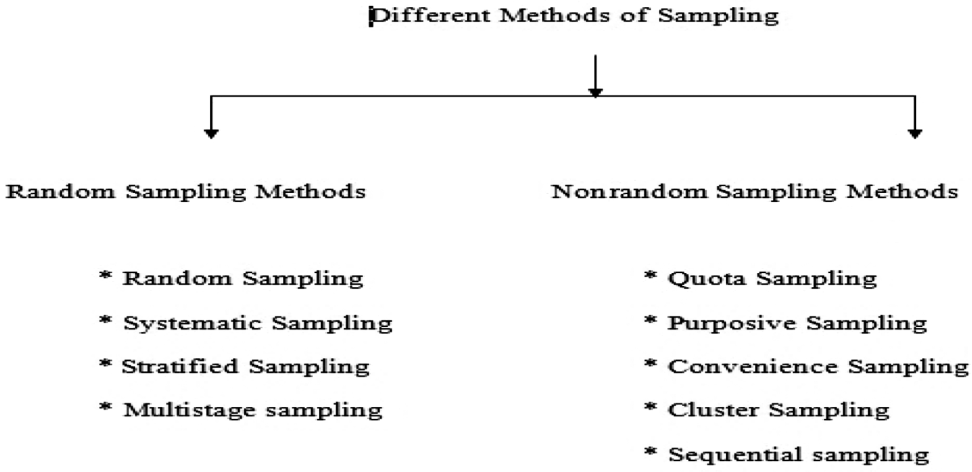
Normally, the sampling is based on 2 specific principles.

Principle 1: Law of statistical regularity

This law implies that a reasonably large number of items is selected at random from the population in such a way that the characteristics of the population and the sample are equal.

Principle 2: Law of inertia of large numbers

This law reveals that wherever the sample is quite large, the inference will be close to the actual.



Different Methods of Sampling

Random Sampling Methods

- Random Sampling
- Systematic Sampling
- Stratified Sampling
- Multistage sampling

Nonrandom Sampling Methods

- Quota Sampling
- Purposive Sampling
- Convenience Sampling
- Cluster Sampling
- Sequential sampling

12.10 Random Sampling

According to N. M. Harper, '[random sampling] is a sample selected in such a way that every item in the population has an equal chance of being included'. In general, it is the process of selecting sample from a population in such a way that every item of the population has an equal chance of being included in the sample.

Example:
Selection of any 5 members out of a group containing 20 members will constitute a random sample.

Example:

Selection of 4 aces out of a well-shuffled pack of 52 cards will constitute a random sample.

Notations:

Population size:	N
Sample size:	n ($n \leq N$)
Number of Possible samples:	$m = {}^{N}C_{n}$
Different samples:	$S_1, S_2, \ldots, S_m$
P(Selecting a sample)=	$1/m$

In other words, simple random sampling refers to the process that ascertains that each sample of size n ($S_1, S_2, \ldots, S_m$) has an equal probability of being selected up of the chosen sample.

The simple random sampling method can be adopted with or without replacement of the items selected. In practice, sampling is done always without replacement. While selecting a single random sample, we must use some specific method to ensure true randomness. One such method involves the use of random number. Using a random numbers ensures that every element in the population has an equal and independent chance of being selected.

Example:

Let us consider the production record on a particular day of the employees of a firm BHAVANA SREE LTD. Along with the employee number.

E.No.	Prod.	E.No.	Prod.	E.No.	Prod.	E.No.	Prod.	E.No.	Prod.
001	30	011	59	021	65	031	47	041	55
002	38	012	56	022	42	032	64	042	32
003	33	013	65	023	73	033	55	043	31
004	49	014	50	024	44	034	50	044	35
005	33	015	54	025	54	035	65	045	36
006	43	016	61	026	67	036	53	046	59
007	60	017	71	027	49	037	32	047	68
008	31	018	57	028	38	038	44	048	26
009	34	019	26	029	59	039	38	049	72
010	61	020	41	030	42	040	37	050	29

E. No., Employee number; Prod., Production.

We can use the random number table for selecting a simple random sample of size 5 without replacement from the population of 50 employees.

Step 1: Select 5 two-digit random number using the random number table

04	10	37	17	50

Step 2: Select the employees by considering the random number selected as their employee number.

Random Numbers	04	10	37	17	50
Sequence in Sample	1	2	3	4	5
Production record	49	61	32	71	29

If we proceed in the same way, we can create different samples of size 5.

NOTE: Because we are sampling without replacement, we do not want to use the same random number twice.

12.10.1 Systematic Sampling

Systematic sampling is a procedure that starts with a random starting point in the population and then includes the sample every be *k*th element encountered thereafter.

Example:

Population size (N) : 100 students

Sample size (n) : 10 students

Sampling ratio = $n/N = 10/100 = 1/10$

Form 10 different groups according to roll numbers

G1	:	1	2	3	4	5	6	7	8	9	10
G2	:	11	12	13	14	15	16	17	18	19	20
…		…	…	…	…	…	…	…	…	…	…
G10	:	91	92	93	94	95	96	97	98	99	100

Select any one number in G1

[1 2 3 4 5 6 7 8 9 10]

Suppose the selected item is 8. Then in each group one has to select the 8th item.

That is 8, 18, 28, and 98. The collection of all these elements leads to a sample of size 10. This sample is known as 'systematic sample'.

It is different from the simple random sampling. In this, only the first element is selected randomly. There is a chance of response bias to occur. This method of selecting a sample is commonly used among the probability sampling deigns.

12.10.2 Stratified Sampling (P, N)

(P_1, N_1)	(P_2, N_2)	(P_3, N_3)
(S_1, n_1)	(S_2, n_2)	(S_3, n_3)

P : Population (size N)

P_1, P_2, P_3 : Sub-population (size N_1, N_2, N_3 and $N = N_1 + N_2 + N_3$)

S_1, S_2, S_3 : Samples from each sub-population of size n_1, n_2, and n_3, respectively.

Dividing the single population into many sub-population is called *strata*. Select a random sample from each stratum. Then the stratified sample is the grouping of different sample selected from all the strata with a one sample. This sampling technique needs prior knowledge about the population. This helps to partition the single population into different strata based on some homogeneous characteristics.

To set the maximum information using stratified sampling, the strata must be different from each other but homogeneous within each structure.

Example:

Problem: Determining the faculty preferences for a union in a college

Population: 100

College

To say specifically, the preferences will be differing according to the different grades of the teachers. If we take a sample out of this population directly, we won't get any fruitful results. Instead, try to split this single population of college teachers into different sub-populations based on their grades, and select a sample from each strata and form a one big sample by merging all the sub-samples collected from different strata. If so, there is more chance for us to have fruitful results.

Population: 100

Professors	Assistant Professors	Associate Professors/Readers	Lecturer SG	Lecturer Sr. SG 25 (S_5)
05	5	30	15	Lecturer (S_6)
(S_1)	(S_2)	(S_3)	(S_4)	20

$$\text{Stratified sample} = (S_1)U(S_2)U(S_3)U(S_4)U(S_5)U(S_6)$$

In stratified sampling, the number of items selected from each stratum is in proportion to its size. This method ensures that the stratum in the sample is overweighted by the number of elements it contains in it. It is used in managerial applications because it allows conclusions to be inferred based on each stratum separately.

12.10.3 Multistage Sampling

As the name indicates, the selection process of this type of sample contains different stages.

Stage 1: Population is divided into different groups called 'first-stage units'.

Stage 2: The first-stage units are then divided into smaller groups, called 'second-stage units'.

Stage 3: The second-stage units are divided into smaller groups, called 'third-stage units'.

This staging process will go on until a sample of required number is attained.

Example: Population: Group of institutions

I_1	I_2	I_3
I_4	I_5	I_6

I: each institution contained with different department.

I_1

D_1	D_2	D_3
D_4	D_5	D_6

D: each department contain different courses

D_1

C_1	C_2	C_3

First-stage units: $[I_1, I_2, \ldots, I_6]$

Second-stage units: $[I_1(D_1, D_2, \ldots, D_6), \ldots]$

Third-stage units: $[(I_1, D_1)(C_1, C_2, C_3), \ldots]$

Select a sample using proper methods out of first-stage units. Then select a sample out of second-stage units selected based on first-stage units, and the same procedure is repeated from stage to stage until the required sample size is reached. This method of selecting sample will be useful in large populations.

12.11 Nonrandom Sampling Methods

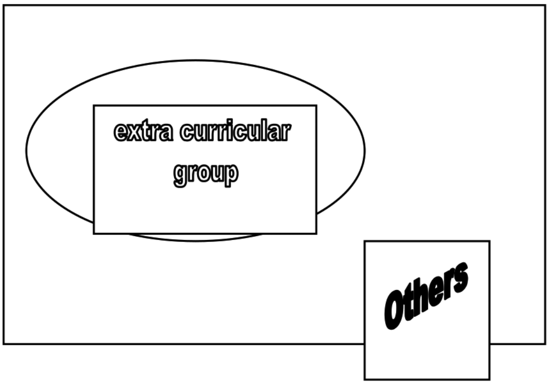
To apply the probability, sampling needs a list of all sampling units. The same is not possible in all cases. To overcome this situation, we seek the help of nonrandom sampling technique.

12.11.1 Convenience Sampling

In this type of sampling, the selection of sample is totally left to the convenience of the researcher. The cost of selecting a convenience sample is low compared with the probability sampling. On the other hand, it suffers from excessive biasness, which leads to possible errors and it cannot be quantifiable. It is useful in public opinion surveys, samples regarding demand analysis, shopping centre surveys, and so on. Convenience sampling is separately used in exploratory studies or when representing the population is not a critical factor.

12.11.2 Purposive Sampling

If we select an element from the population based on certain characteristics, then the resulting sampling is known as 'purposive' or 'judgment sample'.



Population of Students
Among the 100 students in a class, the sample is selected based only on the students those who are members of extracurricular group.

12.11.3 Quota Sampling

When there is a defined proportion of an element to be selected from the population based on certain characteristics, it is called 'quota sampling'.

Example:

Population: 1000 Customers	
Top Income Group	(TIG) 20%
Middle Income Group	(MIG) 30%
Low income (LIG)	Group 50%

Out of this population select a sample of size 100, in such a way that

Sample: 100 Customers	
Top Income Group	(TIG) 30%
Middle Income Group	(MIG) 30%
Low income (LIG)	Group 40%

This type of sampling is often used in conducting public opinion polls such as predicting consumer preferences in market research studies and public opinions regarding political issues and candidates. There is a chance of reducing the biasness in the case. It is easy to adopt and less costly.

12.11.4 Cluster Sampling

Cluster sampling requires the prior knowledge about the population. The population is to be partitioned into different groups called 'clusters'; the formation of clusters is based on some characteristic.

Step 1: Form the clusters

Step 2: Select few clusters at random

Step 3: Select the elements at random based on the randomly selected clusters

The resulting sample is known as 'cluster sampling'.

Example:

Population: 1000 students

Clusters formed based on discipline

Department of Mathematics, 50	Department of Computer Science, 100	Department of Management, 500
Department of Fashion, 150	Department of Bio-Tech, 50	Department of Interior Design, 150

Among the clusters randomly select any 2 clusters.

Department of Fashion, 50	Department Computer Science, 100

Select few elements randomly out of these 2 randomly selected clusters.

Department of Fashion, 5	Computer Science, 15

The above sample is said to be a cluster sample of size 20.

12.11.5 Sequential Sampling

Samples are selected one after another based on the outcome of the previous samples.

This type of sampling method is used in the statistical quality control department often.

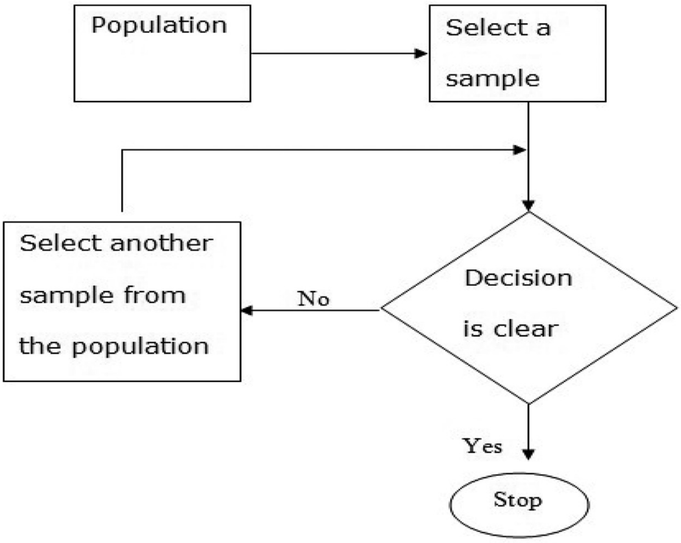
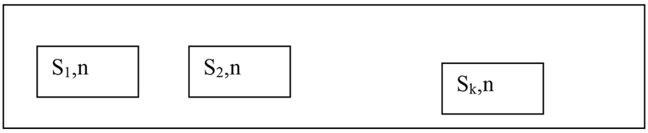


FIGURE 12.1
Flowchart referring the way of selecting the sample from the population.

12.12 Sampling Distributions

We can define a sampling distribution as shown here. The distribution of all possible values that can be assumed by some statistic evaluated from samples of the same size randomly drawn from some population is called the 'sampling distribution' of that statistic.

Population: N



From the population of size N, draw the different sample of size n, $(n < N)$ randomly. Let the sample be (s_1, n), (s_2, n), … (s_k, n). [Refer figure 12.2].

With the sample data, it is possible to evaluate the sample statistics like sample mean, sample SD, and etc.

Sampling Distribution Based on the Sample Means:

Consider all the sample means $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$.

Construct a frequency distribution based on the means of the samples.

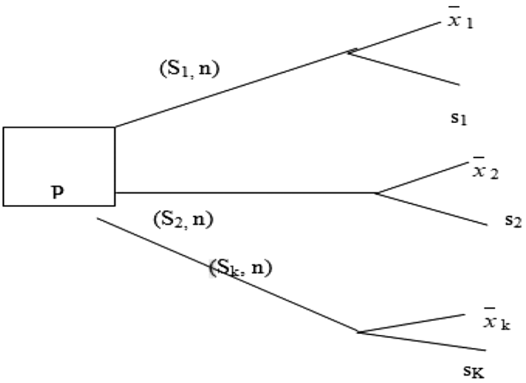


FIGURE 12.2
Tree diagram based on the *n* – samples.

Means of Sample	Frequency

The resulting continuous distribution based on the means of the sample is defined as sampling distribution based on the means of the samples. For the constructed distribution, it is possible for us to evaluate the measures mean, SD, etc.

The mean is said to be the mean of the sample means. The standard deviation of this sampling distribution based on the mean is known as the standard error (SE) of the distribution.

In the same way, one can construct a sampling distribution based on the SD of the samples.

SDs of Sample	Frequency

Likewise, for every statistic of the sample, it is possible to construct different sampling distribution.

Example:

Population: Weekly expense of 5 families

Family	1	2	3	4	5
Expense (Rs)	45	40	47	35	33

Collect all possible combinations of different samples containing exactly of size 2. Also evaluate the sample means and SDs, as well as, the mean and SD of the population. Because $N = 5$ and $n = 2$, we can have ${}^5C_2$ samples. Overall, we can have 10 sample of size 2.

Sample No.	Sample Data	Sample Mean
01	45,40	42.5
02	45,47	46.0
03	45,35	40.0
04	45,33	39.0
05	40,47	43.5
06	40,35	37.5
07	40,33	36.5
08	47,35	41.0
09	47,33	40.0
10	35,33	34.0
Total		400

Construction of a sampling distribution

Mean of the population = 40

SD of the population = 5.44

Consider all the sample means and the associated sampling distribution of $\bar{x}$ is

$\bar{x}$	Frequency	P($\bar{x}$)	$\bar{x} - \mu$	$(\bar{x} - \mu)^2$
46	1	1/10	6	36
43.5	1	1/10	3.5	12.25
42.5	1	1/10	2.5	6.25
41	1	1/10	1	1
40	2	2/10	0	0
39	1	1/10	–1	1
37.5	1	1/10	–2.5	6.25
36.5	1	1/10	–3.5	12.25
34	1	1/10	–6	36

We now evaluate E($\bar{x}$) and var($\bar{x}$)

$$\text{E}(\bar{x}) = \sum_{i=1}^{10} [p(x_i) \times x_i]$$

$$= (1/10)\ [46] + 1/10\ [43.5] + \ldots + 34\ [1/10] = 40$$

$$\text{Var}(\bar{x}) = \text{E}\left(\left(x - \mu\right)^2\right) = \sum_{i=1}^{10} [p(x_i) \times (x_i - \mu)^2]$$

$$= (1/10)(36) + (1/10)\ (12.25) + \ldots + (1/10)\ (36)$$

$$\text{Var}(\bar{x}) = 11.11;\ \sigma_{\bar{x}} = 3.331$$

$$\text{Var}(\bar{x}) = \sqrt{\frac{\sigma^2}{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{29.59}{2}}\sqrt{\frac{5-2}{5-1}} = 11.1$$

$$\sigma_{\bar{x}} = 3.331$$

12.13 Need for Sampling Distribution

We can draw the inferences about the population parameters based on the sample statistics only. In addition to the sample statistic, if we know the probability distributions with respect to the sample statistic, it is possible for us to calculate the probability when the sample statistic assumes any specific value. This characteristic is needed in all statistical inferences.

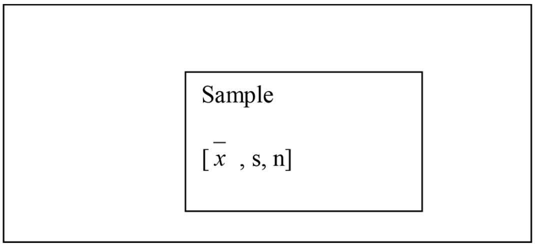
NOTE: The variance of the sampling distribution is equal to the variance of the population divided by the size of the sample used to get the sampling distribution.

Case 1: ${\sigma_{\bar{x}}}^2 = \frac{\sigma^2}{n}$; when the population size is infinite
Case 2: ${\sigma_{\bar{x}}}^2 = \frac{\sigma^2}{n}\left[\frac{N-n}{n-1}\right]$; when the population size is finite

Central Limit Theorem

P: (μ, σ, N) for a sufficiently large value of n ($n \geq 30$), the sampling distribution of sample mean ($\bar{x}$) is approximately a normal distribution with mean μ and $\sigma_{\bar{x}}$.

P: (μ, σ, N)



NOTE: The same holds good for the sample proportion also.

Relationship between the sample statistics with the population parameter

- The mean of all possible sample means will be exactly equal to the universe mean.
- The mean of all possible sample SDs ($\sigma_{\bar{x}}$) will be approximately equal to $\frac{\sigma}{\sqrt{n}}$; where n is the sample size.

NOTE: While evaluating the sample variance, we use the relation.

$$s^2 = \frac{\sum_{i=1}^{n}[xi - \bar{x}]^2}{n-1}$$

Here we use (n – 1) in the division instead of (n).

This is due to a technical reason to have E(s^2) = σ²

Show that the sample variance $s^2 = \frac{\sum_{i=1}^{n}[x_i - \bar{x}]^2}{n-1}$ is an unbiased estimator of the population variance σ².

Case 1:

Sample from an infinite population having normal distribution, we know that the expected value of the chi-square statistic $\frac{[n-1]s^2}{\sigma^2}$ is [n – 1]

$$\text{that is, } E\left[\frac{(n-1)s^2}{\sigma^2}\right] = n-1; \frac{(n-1)}{\sigma^2}\,\mathrm{E}(s^2) = n-1$$

This implies that, E(s^2) = σ²

The sample variance s^2 is an unbiased estimate of σ² for infinite populations having normal distributions.

Case 2:

For samples from infinite populations.

$$s^2 = \frac{\sum_{i=1}^{n}[x_i - \bar{x}]^2}{n-1} \tag{12.1}$$

Taking the expectation on both sides of (12.1) we have

$$\mathrm{E}[s^2] = \mathrm{E}\left\{\frac{\sum_{i=1}^{n}[x_i - \bar{x}]^2}{n-1}\right\} = \frac{1}{n-1}E\left[\sum_{i=1}^{n}\{[x_i - \mu] - [\bar{x} - \mu]\}^2\right]$$

it is obvious E [(x_i – μ)²] = σ²

$$\sigma_{\bar{x}}{}^2 = \mathrm{E}\,((\bar{x} - \mu)^2) = \sigma^2/\mathrm{n}$$

$$\mathrm{E}\,(\mathrm{S}^2) = \frac{1}{n-1}\left\{E\left[\sum_{i=1}^{n}\{[x_i - \mu]^2] - n[\bar{x} - \mu]\}^2]\right]\right\} = \frac{1}{n-1}\left\{\left[\sum_{i=1}^{n}\{E[x_i - \mu]^2] - n[\bar{x} - \mu]\}^2\right]\right\}$$

$$= \frac{1}{n-1}\left\{\left[\sum_{i=1}^{n}\left\{\left[Ex_i - \mu^2\right] - n\frac{\sigma^2}{n}\right\}\right]\right\} = \frac{1}{n-1}\left\{\left[\sum_{i=1}^{n}\sigma^2 - n\frac{\sigma^2}{n}\right]\right\} = \frac{1}{n-1}\{n\sigma^2 - \sigma^2\} = \sigma^2$$

$$\mathrm{E}\,(\mathrm{S}^2) = \sigma^2$$

The sample variance is thus an unbiased estimate of σ² for an infinite population in general.

12.14 Standard Error for Different Situations

12.14.1 When the Population Size Is Infinite

1. Standard error (SE) of the specified sample mean $\bar{x}$.

$$\text{SE} = \text{SE}[\bar{x}] = \left[\frac{\sigma}{\sqrt{n}}\right]; \text{ if } \sigma \text{ is known}$$

$$\text{otherwise SE}[\bar{x}] = \frac{s}{\sqrt{n}}$$

2. Standard error (SE) of difference of 2 sample means $[\bar{x}_1 - \bar{x}_2]$

$$\text{SE} = \text{SE } [\bar{x}_1 - \bar{x}_2] = \sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}; \text{ if both the population SDs are known}$$

$$\text{Otherwise, SE } [\bar{x}_1 - \bar{x}_2] = \sqrt{\frac{s_1{}^2}{n_1} + \frac{s_2{}^2}{n_2}};$$

3. Standard error (SE) of the specified sample SD(s)

$$\text{SE} = \text{SE (s)} = \left[\frac{\sigma}{\sqrt{2n}}\right]; \text{ if } \sigma \text{ is known}$$

$$\text{Otherwise SE (s)} = \left[\frac{s}{\sqrt{2n}}\right];$$

4. Standard error of the difference of two sample SDs s$_1$

$$\text{SE} = \text{SE}(s_1 - s_2) = \sqrt{\frac{\sigma_1{}^2}{2n_1} + \frac{\sigma_2{}^2}{2n_2}}; \text{ if } \sigma_1 \text{ and } \sigma_2 \text{ are known.}$$

$$\text{Otherwise SE}(s_1 - s_2) = \sqrt{\frac{s_1{}^2}{2n_1} + \frac{s_2{}^2}{2n_2}};$$

5. Standard error (SE) of the specified sample proportion (p):

$$\text{SE} = \text{SE(p)} = \sqrt{\frac{PQ}{n}}; \text{ if } P \text{ is known, } Q = 1 - P$$

$$\text{Otherwise SE(p)} = \sqrt{\frac{pq}{n}}; \text{ if } P \text{ is known}$$

6. Standard error (SE) of the difference of two sample proportions

$[P_1 - P_2]$

$$\text{SE} = \text{SE } (P_1 - P_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}; \text{ if } P_1 \text{ and } P_2 \text{ are known}$$

$$\text{Otherwise SE } (P_1 - P_2) = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$$

7. Standard error [SE] of the sample correlation coefficient [*r*]

$$\text{SE} = \text{SE}(r) = \left[\frac{1-R^2}{\sqrt{n}}\right]; \text{ if } R \text{ is known}$$

8. Otherwise SE(*r*) = $\left[\frac{1-r^2}{\sqrt{n}}\right]$

12.14.2 When the Population Is Finite

Sample is drawn with replacement

1. Standard error of the specified sample mean ($\bar{x}$), use formula [1]
2. Standard error of the specified sample proportion (p), use formula [5].

Sample is drawn without replacement

1. Standard error [SE] of the specified sample mean ($\bar{x}$):

$$\text{SE} = \text{SE}(\bar{x}) = \left[\frac{\sigma}{\sqrt{n}}\right]\left[\sqrt{\frac{N-n}{N-1}}\right]; \text{ if } \sigma \text{ is known}$$

$$\text{Otherwise SE} = \text{SE}(\bar{x}) = \left[\frac{s}{\sqrt{n}}\right]\left[\sqrt{\frac{N-n}{N-1}}\right]$$

2. Standard error of the specified sample proportion (*P*):

$$\text{SE} = \text{SE}(p) = \sqrt{\frac{PQ}{n}}\left[\sqrt{\frac{N-n}{N-1}}\right]; \text{ if } P \text{ is known}$$

$$\text{Otherwise SE}(p) = \sqrt{\frac{pq}{n}}\left[\sqrt{\frac{N-n}{N-1}}\right]$$

12.14.3 Sampling Distribution Based on Sample Means

Consider a random sample of size *n* out of a population with actual mean $\bar{x}$ and variance σ^2, then we know that the sample observations are independent and identically distributed random variables. Then, the sample mean, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Clearly $\bar{x}$ is also a random variable with an expected value.

$$\text{E}[\bar{x}] = \text{E}\left\{\frac{\sum_{i=1}^{n} x_i}{n}\right\} = [1/\text{n}]\ \text{E}\left\{\sum_{i=1}^{n} x_i\right\} = [1/\text{n}] \sum_{i=1}^{n} E[x_i] = [1/\text{n}]\{\text{n}\mu\} = \mu$$

$$\text{E}[\bar{x}] = \mu$$

Variance of $\bar{x}$ can be given as

$$\text{Var}[\bar{x}] = \text{Var}\left\{\frac{\sum_{i=1}^{n} x_i}{n}\right\} = \text{Var}\left\{[1/\text{n}]\sum_{i=1}^{n} x_i\right\} = [1/\text{n}^2]\left\{\sum_{i=1}^{n} \text{var}[x_i]\right\} = [1/\text{n}^2][\text{n}\sigma^2] = \sigma^2/\text{n}$$

$$\text{Var}[\bar{x}] = \sigma^2/\text{n};\ \text{SD}[\bar{x}] = \frac{\sigma}{\sqrt{n}}$$

NOTE: It indicates that the expected value of the sample mean and the actual population mean are one and the same. This shows that the variability in sample means is lesser than the population variance, $\underset{n\to\infty}{Lt}\ [\sigma_{\bar{x}}] = 0$. Whenever the sample size is large, the fluctuation will be less from one sample to the other.

Population parameters are estimated from sample data because it is not possible to examine the entire populations practically to make a perfect evaluation.

Statistical estimation procedures provide the process by which estimates of the population parameters can be evaluated by the degree of confidence needed. This degree of confidence is controllable with respect to the size of the sample and by the type of estimate made.

12.15 Point and Internal Estimation

Refer to Figure 12.3

Type of Organization	Estimation of Interest
Manufacturing industry	Quality of raw materials used for production
Bank	Mean number of arrivals of the customer at the teller's window

The estimate can be of 2 types, they are

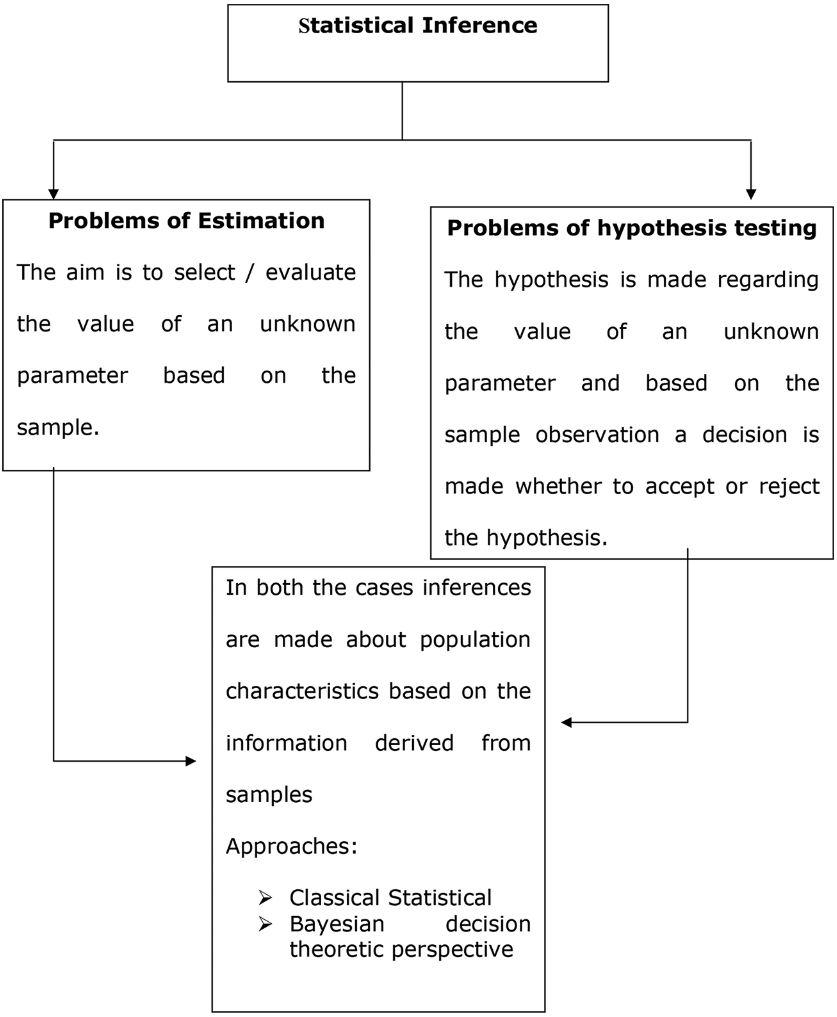
1. point estimates and
2. interval estimates

12.15.1 Point Estimate

A point estimate refers a specific value used to estimate the value of the unknown population parameter.

Example

- The mean salary of a sample of top-level executives in many firms may be used as a point estimate of the corresponding population mean for top-level executives in all firms.
- The percentage of employed women who prefer Cinthal brand soap over all other brands may be used as an estimate of the corresponding population percentage of all employed women.



Statistical Inference

Problems of Estimation

The aim is to select / evaluate the value of an unknown parameter based on the sample.

Problems of hypothesis testing

The hypothesis is made regarding the value of an unknown parameter and based on the sample observation a decision is made whether to accept or reject the hypothesis.

In both the cases inferences are made about population characteristics based on the information derived from samples

Approaches:

- Classical Statistical
- Bayesian decision theoretic perspective

FIGURE 12.3
Point and internal estimation.

Similarly, one can use the sample mean to estimate the population mean, the use of sample SD to estimate the population S, and so on; in each case, we use point estimate of the parameter.

Estimate and Estimator

An estimator is a random variable, and its numerical value is an estimate.

Population Parameter	Estimator (Sample Statistic)	Estimate (Value of Estimator)
Mean – μ	$\overline{x}$	$\overline{x} = 100$
Variance σ^2	s^2	$s^2 = 50$

12.15.2 Properties of Good Point Estimators

The criteria for good point estimators are

1. unbiased,
2. relative efficiency,
3. consistency, and
4. sufficiency.

Unbiased

An estimator is unbiased, if its expected value is equal to the population parameter being estimated.

Relative Efficiency

The sampling variability of an estimator is known as 'relative efficiency'.

If 2 estimators of a given population parameter are both unbiased, the one with the smaller variance for a given sample size is defined as being relatively more efficient. If e_1 and e_2 are 2 unbiased estimators of the parameter e, then the relative efficiency of e_1, with respect to e_2 is defined as (assume that var (e_1) < var (e_2))

$$\text{Relative efficiency} = \frac{Var[e_2]}{Var[e_1]}$$

Consistency

An estimator is said to be consistent, if the probability of the parameter being estimated approaches 1 as n approaches infinity.

$$\text{That is, Lim (P } (e_1 - e) < \epsilon) = 1$$

$$n \to \infty$$

e_1: sample estimator
e: population estimator

Sufficiency

An estimator, e_1, is said to be a sufficient estimator if it uses all the information contained in the sample to estimate the population parameter.

12.16 Interval Estimate

An interval estimate of a population parameter is the specification of 2 values between which we have a certain degree of confidence that actual population parameter lies. It can be otherwise called a 'confidence internal estimation'. To evaluate the same, we require the value for the confidential level or the level of significance.

Population parameter : μ

Sample parameter : $\bar{x}$, s, n

Level of significance : 5%

Test statistic : Z

Table value of the test statistic: Z_t

$$Z_{0.05} = 1.96 \text{ (2-tailed test)}$$

Then the interval estimation of the population parameter μ can be defined as μ: $\bar{x} \pm Z_t * SE[\bar{x}]$; where $SE[\bar{x}] = \frac{s}{\sqrt{n}}$

$$\text{Then } \mu\text{: } \bar{x} \pm [Z_t] * \frac{s}{\sqrt{n}}$$

$$\mu\text{: } \bar{x} \pm [1.96] * \frac{s}{\sqrt{n}}$$

There is a 95% confidential level for the population parameter μ to lie in the interval

$$\left[\bar{x} - [1.96] * \frac{s}{\sqrt{n}},\ \bar{x} + [1.96] * \frac{s}{\sqrt{n}}\right]$$

This clearly indicates that there is a 5% chance for the population mean μ not to lie in the defined internal estimate.

12.17 Confidence Interval Estimation for Large Samples

For business application, it is not sufficient merely to consider the single point estimate of the population parameter. Instead, we require an estimation procedure that permits some error in the estimate with the given level of accuracy. In classical inference, such a method incorporates the use of what is known as a 'confidence interval estimation'. We can discuss the same with respect to the population mean as the parameter of interest.

Consider the sampling distribution of $\bar{x}$ (mean) of the random samples of size, n. From a normal population with mean μ and known variance σ^2, that is $N(\mu, \sigma^2)$ the same can be defined in the standard form as, transferred with respect to the Z statistic.

$$Z = \frac{\bar{x} - \mu}{\left\{\frac{\sigma}{\sqrt{n}}\right\}}; \text{ where } Z \sim (0, 1)$$

If we permit the error percentage as α, we say the level of significance is α.

We can assert with the probability (1 – α) that normal random variable

$$Z = \frac{\bar{x} - \mu}{\left\{\frac{\sigma}{\sqrt{n}}\right\}} \text{ will lie in between } -Z\alpha \text{ and } + Z\alpha.$$

The same can be written symbolically, P ($-Z\alpha < Z < +Z\alpha$) = $1 - \alpha$

$$P\left(-Z\alpha < \frac{\bar{x}-\mu}{\left\{\frac{\sigma}{\sqrt{n}}\right\}} < Z\alpha\right) = 1-\alpha$$

$$P\left(-Z\alpha * \left\{\frac{\sigma}{\sqrt{n}}\right\} < \bar{x}-\mu < Z\alpha * \left\{\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha\right.$$

$$P\left(x - Z\alpha * \left\{\frac{\sigma}{\sqrt{n}}\right\} < \mu < \bar{x} + Z\alpha * \left\{\frac{\sigma}{\sqrt{n}}\right\}\right) = 1-\alpha \tag{12.2}$$

Equation (12.2) reveals that μ is contained in the interval between

$$\left(\bar{x} - Z\alpha * \left\{\frac{\sigma}{\sqrt{n}}\right\}, \bar{x} + Z\alpha \times \left\{\frac{\sigma}{\sqrt{n}}\right\}\right) \text{ and its probability equal to } (1-\alpha).$$

The interval ($\bar{x} - Z\alpha \times \left\{\frac{\sigma}{\sqrt{n}}\right\}, \bar{x} + Z\alpha \times \left\{\frac{\sigma}{\sqrt{n}}\right\}$) is referred to as the confidential interval for μ, and ($1 - \alpha$) is called the 'degree of confidence' because μ is contained in the given interval with a probability value ($1 - \alpha$).

Hence, the probability of the value of μ to lie in the interval

$$\left(\bar{x} - Z\alpha \times \left\{\frac{\sigma}{\sqrt{n}}\right\}, \bar{x} + Z\alpha \times \left\{\frac{\sigma}{\sqrt{n}}\right\}\right) \text{ is } (1-\alpha).$$

NOTE: If the sample size is large enough say $n \geq 30$, then the sample is said to be a large sample. If not, it is referred to as a small sample [$n < 30$].

Example:

Suppose we wish to estimate the average weekly sales for Sree Balaji Food store in Trichy Corporation. Also, suppose we know that the population SD of average weekly sales is $4.5. A sample of 50 stores is taken, yielding a mean, *x*, of $70. Determine a 95% confidence interval for the average weekly sales of Sree Balaji Food Stores.

Step 1: $\alpha = 0.05; \sigma = 4.5; n = 50; \bar{x} = 70$

Since, $n = 50 > 30$; it refers a large sample.

According to the standard normal table when $\alpha = 0.05$, the value of $Z\alpha = Z_{0.05} = 1.96$.

Step 2: The interval estimation can be given as $\bar{x} \pm Z_t * SE[\bar{x}]$.

Step 3: $SE[\bar{x}] = \left\{\frac{\sigma}{\sqrt{n}}\right\} = 4.5/\sqrt{50} = 0.6364$

Step 4: Use the value for $\bar{x}$, $Z\alpha$ and $SE[\bar{x}]$, we have

$$\mu: 70 \pm 1.96\ (0.6364);\ \mu: 70 \pm 1.2473$$

The required confidence interval of estimation with 95% confidence level for the average weekly sales of Sree Balaji Food Stores is μ: (68.7527, 71.2473).

Again the interpretation of this confidence interval is that μ will be contained in that interval calculated for 95% of the sample drawn from the stated population. Also, 5% of the time these intervals will not contain μ.

NOTE: There is a very close association between the lengths of interval in which μ lies and α, the level of significance. Whenever α decreases, the length of the interval where in μ lies also increases.

If we want to increase the chance of the value of μ to lie in the estimated interval, try to choose α minimum.

Suppose for the above problem, if we assure the value of $\alpha = 0$.

We have $Z\alpha = Z_0 = 3$.

Hence, the interval estimation becomes,

$$\mu: 70 \pm 3 \times 0.6364; \quad \mu: 70 \pm (1.9092); \quad \mu: (68.0908, 71.9092)$$

Since $\alpha = 0$; There is a 99.73% assured chance for the population mean μ to lie in the interval (68.0908, 71.9092).

NOTE: It is obvious that in the problem, the interval estimation when $\alpha = 0.05$ lies well within the interval estimation when $\alpha = 0$. Also when σ is not known, we can make use of the sample SD(s). Then the interval estimation formula is reduced to $\bar{x} \pm Z_\alpha \times (s/\sqrt{n})$.

Confidence limits for μ, $(\mu_1 - \mu_2)$, P, and $(P_1 - P_2)$ for large random sample

Particulars	95% CL α = 5%	99% CL α = 21%	99.73% CL α = 0.27%
*Population mean, μ	$\bar{x} \pm [1.96] \times SE[\bar{x}]$	$\bar{x} \pm [2.58] \times SE[\bar{x}]$	$\bar{x} \pm [3] \times SE[\bar{x}]$
*Difference between the 2 population means, μ_1 and μ_2	$\{\bar{x}_1 - \bar{x}_2\} \pm$ $[1.96] \times SE[\bar{x}_1 - \bar{x}_2]$	$\{\bar{x}_1 - \bar{x}_2\} \pm$ $[2.58] \times SE[\bar{x}_1 - \bar{x}_2]$	$\{\bar{x}_1 - \bar{x}_2\} \pm$ $3 \times SE[\bar{x}_1 - \bar{x}_2]$
*Population proportion, P	$p \pm [1.96] \times SE[p]$	$p \pm [2.58] \times SE[p]$	$p \pm [3] \times SE[p]$
Difference between the 2 population proportion $P_1 - P_2$	$[p_1 - p_2] \pm$ $[1.96] \times SE[p_1 - p_2]$	$[p_1 - p_2] \pm$ $[2.58] \times SE[p_1 - p_2]$	$[p_1 - p_2] \pm$ $[3] \times SE[p_1 - p_2]$

SE, standard error; CL, confidence limits; α, 10%; $Z_{0.1}$, 1.645.

Example:

In a random sample of size 100 taken from a population of size 1000, the mean and SD of a sample characteristic is found to be 4.8 and 1.1, respectively. Find the 95% confidence interval for the population mean.

Step 1: $\alpha = 0.05$; $s = 1.1$; $n = 100$; $\bar{x} = 4.8$

Since, $n = 100 > 30$; it refers to a large sample.

According to the standard normal table when $\alpha = 0.05$, the value of $Z_\alpha = Z_{0.05} = 1.96$

Step 2: The interval estimation can be given as $\bar{x} \pm Z_\alpha \times SE[\bar{x}]$.

Step 3: $SE[\bar{x}] = \left\{\frac{s}{\sqrt{n}}\right\}$ = 1.1/√100 = 0.11; since the σ value is not known.

Step 4: Use the value for $\bar{x}$, Z_α & $SE[\bar{x}]$, we have

$$\mu: 4.8 \pm 1.96 \times 0.11;\ \mu: 4.8 \pm 0.2156;\ \mu: (4.5844, 5.0156)$$

Step 5:

The required confidence interval of estimation with 95% confidence level is μ: (4.5844, 5.0156)

12.18 Confidence Intervals for Difference between Means

Example:

Bhavana Sree Ltd.'s (BSL) current packaging machinery is known to pour ground coffee into 1-kg cans with an SD of 0.6 grams. BSL is considering a new packaging machine that is said to pour coffee into 1-kg cans with a SD of 0.3 grams. Both the machines pour ground coffee according to a normal distribution. Before deciding to invest, BSL wishes to evaluate the performance of the new machine against that old machine. A sample was taken on each machine to measure the mean weight of the contents of the 1-kg can yielding the following results:

Sample Uses Old Machine	Sample Using New Machine
$n_1 = 30$	$n_2 = 36$
$\bar{x}_1 = 16.7$ grams	$\bar{x}_2 = 15.8$ grams

Construct a 95% confidence interval for the difference in the average weight of contents poured by the old versus the new machine.

Step1: α = 0.05

Because both the samples are large, the table value of $Z_{0.05} = 1.96$

Population 1	Population 2
Mean = μ_1	Mean = μ_2
SD = σ_1 = 0.6 grams	SD = σ_2 = 0.6 grams

Sample Uses Old Machine	Sample Using New Machine
$n_1=30$	$n_2=36$
$\bar{x}_1 = 16.7$ grams	$\bar{x}_2 = 15.8$ grams

Step 2:

The interval estimation can be given as $\{\bar{x}_1 - \bar{x}_2\} \pm Z_\alpha \times SE[\bar{x}_1 - \bar{x}_2]$

Step 3:

$$SE[\bar{x}_1 - \bar{x}_2] = \sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}} = \sqrt{\frac{0.6^2}{30} + \frac{0.3^2}{36}}$$

$$SE[\bar{x}_1 - \bar{x}_2] = \sqrt{0.012 \ + \ 0.0025} = \sqrt{}0.0145 = 0.1204$$

Use the values of $\bar{x}_1, \bar{x}_2$, Zα, and SE, we have

μ = (16.7 – 15.8) ± 1.96 (0.1204)

μ = 0.9 ± 0.236; μ = (0.664, 1.1359)

Step 4: Thus, 0.664 and 1.1359 are the lower and upper bounds respectively of the 95% confidence interval for $[\bar{x}_1 - \bar{x}_2]$.

12.19 Estimating a Population Proportion

Finite Population

Example:

The central government is interested in evaluating the number of *Fortune 500* manufacturing firms that plan to 'fight inflation' by following certain voluntary wage–price guidelines. A sample of 100 of the firms is taken, and 20 said they did not follow any of these guidelines.

Determine the 90% confidence interval for the percentage of *Fortune 500* firms that do not follow the guidelines.

Step1: α = 0.1

Because the sample is large, the table value of $Z_{0.1} = 1.645$

Sample proportion = $p = \frac{20}{100} = 0.2; q = 1 \ \ p = 0.8; n = 100$

Step 2: The interval estimation of the population proportion can be given as

$$p \pm Z_\alpha \times \text{SE}[p]$$

Step 3:

$$SE(p) = \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{0.2 \times 0.8}{100}}\sqrt{\frac{500-100}{500-1}}$$

$$\text{SE}(p) = 0.04 \times 0.8953 = 0.0358; \text{SE}(p) = 0.0358$$

Step 4: Use the values of *p*, Zα and SE[*p*], we have

P: 0.2 ± 1.645 × 0.0358; P: 0.2 ± 0.0589; P: (0.1411, 0.2589)

Step 5: Thus, 14.11% and 25.89% are the lower and upper bounds, respectively, of the confidence interval.

Example:

A random sample of size 10 is drawn without replacement from a finite population of 30 units. If the number of defective units in the population is 6, find the SE(p).

Step 1: $n = 10$; $N = 30$ (finite population); $P = 6/30 = 1/5 = 0.2$; $Q = 1 - P = 0.8$

Step 2:

$$SE(p) = \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{0.2\times 0.8}{10}}\sqrt{\frac{30-10}{30-1}} = \sqrt{}(0.0110) = 0.105$$

Step 3: The value of SE(P) is 0.105.

12.20 Estimating the Interval Based on the Difference between Two Proportions

Example:

A sample survey of citizens in Trichy showed that among the 1000 members interviewed, 420 would consider the purchase of an electric vehicle, if one were readily available on the market today. In another survey, conducted in Chennai, 370 out of 1000 members interviewed responded similarly. Construct a 99% confidence interval for the true difference in the proportion of favourable responses in the 2 cities.

Step 1:

Sample 1	**Sample 2**
$p_1 = 420/1000 = 0.42$	$p_2 = 370/1000 = 0.37$
$q_1 = 1 - 0.42 = 0.58$	$q_2 = 1 - 0.37 = 0.63$
$n_1 = 1000$	$n_2 = 1000$

$$Z\alpha = Z_{0.01} = 2.58$$

Step 2: $[p_1 - p_2] \pm Z\alpha \times SE[p_1 - p_2]$

Step 3:

$$\text{SE}\left(p_1 - p_2\right) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} = \sqrt{\frac{0.42\times 0.58}{1000} + \frac{0.37\times 0.63}{1000}} = 0.0218$$

$$\text{SE}(p_1 - p_2) = 0.0218$$

Step 4: Use the value of p_1, p_2, $Z\alpha$, and $\text{SE}(p_1 - p_2)$, we have

$$(P_1 - P_2)\text{: } (0.42 - 0.37) \pm 2.58\ (0.0218)$$

$$(P_1 - P_2)\text{: } 0.05 \pm 0.0562$$

$$(P_1 - P_2): (-0.0062, 0.1062)$$

Because the value of probability value is ≥0; we discard the negative value.

Hence, $(P_1 - P_2)$: (0, 0.1062)

Step 5:

Thus, 0 and 0.1062 are the lower and upper bounds respectively, of the 99% confidence interval for $(\mathrm{P}_1 - \mathrm{P}_2)$.

12.21 Confidence Interval Estimation for Small Sample

Example:

In an effort to establish a standard time needed to perform a specific table, a production engineer selects 16 experienced employees to perform the table. The mean time required by the 16 employees is 13 minutes. The SD is 3 minutes. The production engineer wishes to construct a 95% confidence internal for the true mean length of time to complete the table.

Step 1:

Sample

$$\text{Mean} = \bar{x} = 13$$

$$\text{SD} = s = 3$$

$$n = 16$$

$\because n = 16 (< 30)$; it is a small sample. $\alpha = 0.05, df = v = n - 1 = 16 - 1 = 15$.

The table value of $t_t[0.05, 15\ df] = 2.1315$.

NOTE: Because the table value of t is given based on one-tail test, while taking the table value based on two-tail test, consider the value of α as [α/2]. Here α = 0.05, but consider

$$\alpha = 0.025.$$

Step 2: The interval estimation can be given as, $\mu \pm t_\alpha[v] \times SE(\bar{x})$.

Step 3: Find $SE(\bar{x})$

$$SE(\bar{x}) = \frac{s}{\sqrt{n-1}} = \frac{3}{\sqrt{15}} = 0.7746$$

Step 4: Use the values of $\bar{x}, t_\alpha(v)$, and $\mathrm{SE}(\bar{x})$, we have

$$\mu : 13 \pm (2.1315)(0.7746)$$

$$\mu : 13 \pm 1.6511$$

Step 5: The required confidence internal of estimation with 95% confidence level is $\mu : [11.3489, 14.6511]$

Example:

A simple random sample of 16 radio stations is selected to estimate the average charge for the same fixed-length spot announcement. The sample mean and SD are \$15.50 and \$8.00, respectively. Construct the 95% confidence internal for the population mean.

Step 1:

Sample

$$\text{Mean} = \bar{x} = \$15.50$$

$$\text{SD} = s = \$8$$

$$n = 16$$

Because $n < 30$; implies, it refers to a small sample. $\alpha = 0.05, d.f = 16 - 1 = 15$. The table value of

$$t_t[0.05,15 \text{ df}] = 2.131.$$

Step 2: The interval estimation can be given as $\bar{x} \pm t_\alpha(v) \times SE(\bar{x})$

Step 3: Find SE($\bar{x}$)

$$SE(\bar{x}) = \frac{s}{\sqrt{n-1}} = \frac{8}{\sqrt{15}} = 2.0656$$

Step 4: Use the value of $\bar{x}, t_\alpha(v)$, and $SE(\bar{x})$, we have

$$\mu : 15.5 \pm (2.131)(2.0656)$$

$$\mu : 15.5 \pm 4.4018$$

$$\mu : [11.0982, 19.9018]$$

Step 5: The required confidence interval of estimation with 95% confidence level is

$$\mu : [11.0982, 19.9018]$$

Example:

Experimenters test 2 types of fertilizer for possible use in the cultivation of cabbages. They grow cabbages in 2 different fields. One of the fertilizers is applied in each field. At harvest time they select a random sample of 25 cabbages from the crop grown with fertilizer 1. They randomly selected 12 cabbages from the crop grown with fertilizer 2. The sample mean and variance of the weights of cabbages grown with fertilizer 1 are 44.1 g and 36 g. The mean weight computed from the second sample is 31.7 g and the variance is 44 g. The experiments assume that the 2 population weights are normally distributed. They also assume that the 2 population variances are equal. Compute 95% confidence interval for $[\mu_1 - \mu_2]$.

Step 1:

Sample 1	Sample 2
$\bar{x}_1 = 44.1$	$\bar{x}_2 = 31.7$
$\sigma_1^2 = 36$	$\sigma_2^2 = 44$
$n_1 = 25$	$n_2 = 12$

Sample 1 and Sample 2 are small samples.

$$\alpha \; = 0.05; df = [25 - 1] + [12 - 1] = 35; t_\alpha = t_{0.01}\,[35\,df] = 2.0301$$

Step 2: The interval estimation can be given as

$$[\mu_1 - \mu_2]\text{: } \{\overline{x_1} - \overline{x_2}\} \pm t_\alpha \times SE[\overline{x_1} - \overline{x_2}]$$

Step 3:

$$SE[\overline{x_1} - \overline{x_2}] = s_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\text{; where } s_c \text{ can be defined as}$$

$$s_c = \sqrt{\frac{n_1 s_1{}^2 + n_2 s_2{}^2}{n_1 + n_2 - 2}}$$

$$s_c = \sqrt{\frac{n_1 s_1{}^2 + n_2 s_2{}^2}{n_1 + n_2 - 2}} = \sqrt{\frac{25[36] + 12[44]}{25 + 12 - 2}} = 6.3875$$

$$SE[\overline{x_1} - \overline{x_2}] = \mathrm{s_c}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 6.3875\sqrt{\frac{1}{25} + \frac{1}{12}} = 2.2433$$

Use the values of $\bar{x}_1, \bar{x}_2$, t_α and SE, we have

$$[\mu_1 - \mu_2]\text{: } \{\overline{x_1} - \overline{x_2}\} \pm t_\alpha \times SE[\overline{x_1} - \overline{x_2}]$$

$$[\mu_1 - \mu_2]\text{: } \{44.1 - 31.7\} \pm 2.0301 \times 2.2433$$

$$[\mu_1 - \mu_2]\text{: } \{12.4\} \pm 4.5541$$

$$[\mu_1 - \mu_2]\text{: } [7.8459, 16.9541]$$

Hence, the required confidence interval of estimation with 95% confidence level based on difference of 2 means can be given as [7.8459, 16.9541].

Example:

A simple random sample of 10 electronics firms is asked in a questionnaire to state the amount of money spent on employee training program during the year just ended and during a year a decade ago.

Firm	1	2	3	4	5	6	7	8	9	10
Past year, *X*	12	14	8	12	8	10	8	9	10	10
Decade ago, *Y*	10	11	8	7	9	6	10	9	7	9

Construct a 95% confidence interval for the mean difference in expenditures for employee training program by the 10 firms.

Step1: Based on the given data, find the mean difference, $d = x - y$; then find mean and SD based on the values of *d*.

Firm	x	y	$d = x - y$	$d - \bar{d}$	$[d - \bar{d}]^2$
1	12	10	2	0.5	0.25
2	14	11	3	1.5	2.25
3	8	8	0	0	0
4	12	7	5	3.5	12.25
5	8	9	−1	−2.5	6.25
6	10	6	4	2.5	6.25
7	8	10	−2	−3.5	12.25
8	9	9	0	−1.5	2.25
9	10	7	3	1.5	2.25
10	10	9	1	−0.5	0.25
Total			15		44.25

NOTE: We can chose either $[x - y]$ or $[y - x]$ as *d*; provided the sum of *d* is positive.

$$\text{Mean} = \Sigma d/10 = 15/10 = 1.5;\ \text{SD} = s = \{[1/10]\Sigma[d - \bar{d}]^2\}^{[1/2]} = 2.1036$$

$$\alpha = 0.05;\ df = [10-1] = 9;\ t_\alpha = t_{0.05}[9\ df] = 2.262$$

Step 2: The interval estimation can be given as

$$\{\bar{d}\} \pm t_\alpha \times SE[\bar{d}]$$

$$\text{Find } SE(\bar{x}) = \frac{s}{\sqrt{n-1}} = \frac{2.1036}{\sqrt{9}} = 0.7012$$

Step 3:

Use the values of $\{\bar{d}\}, t_\alpha,$ and $SE[\bar{d}],$ we have

$$\mu_d\text{: } 1.5 \pm [2.262] \times [0.7012]$$

$$\mu_d\text{: } [-0.0861, 3.0861]$$

Step 4:

The required confidence interval of estimation with 95% confidence interval with 9 df is μ_d: [– 0.0861, 3.0861]

12.22 Determining the Sample Size

Deciding the proper sample size is an integral part of any sampling study where inferences need to be made.

Maximum Sample Size: Waste of time and money.

Optimum Sample Size: Best

Minimum Sample Size: Accuracy will be lost.

Error

Error is defined as the absolute difference between the parameter being estimated and the point estimate obtained from sample.

Evaluation of sample size for a mean

$$\text{Known elements : } \sigma^2, \bar{x}$$

$$\text{To be estimated : } \mu \sim N(\mu, \sigma^2)$$

The error can be defined as,

$$\text{Error} = |\bar{x} - \mu| \tag{12.3}$$

By definition

$$\text{Error} = Z\alpha \times (\sigma/\sqrt{}\text{n}) \tag{12.4}$$

Equations [12.3] and [12.4] imply that

$$\bar{x} - \mu = Z\alpha \times (\sigma / \sqrt{n}) \tag{12.5}$$

Squaring on both sides of Equation [12.5], we have

$$[\bar{x} - \mu]^2 = [Z\alpha \times (\sigma / \sqrt{n})]^2$$

$$n = \frac{Z_\alpha{}^2 \sigma^2}{[\bar{x} - \mu]^2} \tag{12.6}$$

Thus, Equation [4] gives the sample size required to attain the tolerable error with the required degree of confidence.

NOTE 1: When σ^2 is not known, we can make use of the sample variance s^2 and the sample size n is defined as n = $\frac{t_\alpha{}^2 s^2}{[\bar{x}-\mu]^2}$. The value and it can be referred from the t – table minimum level of significance α and (n – 1) degrees of freedom.

NOTE 2: The sample size for a proportion can be defined as n = $\frac{Z_\alpha{}^2 PQ}{[p-P]^2}$;

When P is not known, it can be assumed that $P = 0.5$.

NOTE 3: For a 2-sample case $[n_1 = n_2 = n]$ the size of the sample can be defined as $n = \frac{Z\alpha^2[\sigma_1^2+\sigma_2^2]}{d^2}$, where d is equal to one-half the width of the desired confidence interval.

NOTE 4: For a 2-sample case, proportions can be defined as $n = \frac{Z\alpha^2[p_1q_1+p_2q_2]}{d^2}$, where d is equal to one-half the width of the desired confidence interval.

Example:

Evaluate the sample size n to find the 90% confidence interval for the purchase price of TVs in various retail stores in a given area such that the sample mean $\bar{x}$ will differ by no more than \$25. Assume that σ is known and equal to \$35/ –.

Step 1:

$$\bar{x} - \mu = 25; \sigma = 35; \alpha = 10\% = 0.1$$

Step 2: $n = \frac{Z\alpha^2\sigma^2}{[\bar{x}-\mu\}^2}$

$$n = \frac{[1.645]^2[35]^2}{[25]^2} = 5.3038; n \geq 5.3038.$$

The sample size should be minimum 6 to attain the error factor 25 with the required 90% confidence level.

Example:

A researcher wishes to know whether the mean length of employment with the current firm at time of retirement is different for men and women. The researcher would like to have a confidence interval estimate of the difference between the population means. The specifications are a confidence interval width of 1 year and 95% confidence. Pilot samples yielded variances of 5 and 7. The researcher wants sample of equal size. What size sample should be drawn from each population?

Step 1: $\alpha = 5\% = 0.05; \sigma_1^2 = 5; \sigma_2^2 = 7; d = ½ = 0.05; Z_\alpha = 1.96$

Step 2: $n \geq \frac{Z_\alpha^2[\sigma_1^2 + \sigma_2^2]}{d^2}$

$$n \geq \frac{[1.96]^2[5+7]}{[0.5]^2} = 184.3968; n \geq 185$$

Step 3: We needed a sample of at least 185 men and an independent sample of at least 185 women is needed.

Example:

A cigarette manufacturer wished to conduct a survey using a random sample to estimate the proportion of smokers who would switch to the company's newly developed low-tar brand. The sampling error should not be more than 0.02 and above or below the actual proportion, with a 99% degree of confidence.

Step 1: $\alpha = 0.01; Z\alpha = 2.58; p - P = 0.02$

Because P is not known, it can be assumed $P = 0.5$

Step 2: $n \geq \frac{Z_\alpha{}^2 PQ}{[p-P]^2}$

$$n \geq \frac{[2.58]^2[.5][.5]}{[.02]^2} = 4160.25; n \geq 4161$$

Hence, the minimum sample size should be at least 4161 members to attain the error 0.02 with the required 99% confidence level.

Example:

The weight of cement bags follows a normal distribution with SD 0.2 kg. Find how large the value of n should be, so that errors can be plus or minus 0.05 of the actual value with a confidence level of 90%.

Step 1: Error = 0.05; $\sigma = 0.2$ kg; $\alpha = 10\% = 0.1$; $Z\alpha = 1.645$

Step 2: Then the value of n can be given as $n \geq \frac{Z_\alpha{}^2 \sigma^2}{[\bar{x} - \mu]^2}$

$$n \geq \frac{[1.645]^2[0.2]^2}{.05^2} = 43.2964; n \geq 43.2964$$

Step 3: The sample size should be at least 44, so that the mean weight of cement bags can be estimated within ± 0.05 Kg of the actual value with a 90% confidence level.

Example:

For 2 populations of consumers, a researcher wants to estimate the difference between the populations who have used a brand of coffee. A confidence coefficient of 0.95 and an interval width of 0.10 are desired. Estimates of p_1 and p_2 are 0.20 and 0.25, respectively. How large should the sample size be ($n_1 = n_2$) ?

Step 1: $\alpha = 0.05; p - P = 0.10$

$$p_1 = 0.2; q_1 = 1 - P_1 = 0.8$$

$$p_2 = 0.25; q_2 = 1 - P_2 = 0.75$$

$$Z\alpha = Z_{0.05} = 1.96$$

Step 2:

$$n \geq \frac{Z_\alpha{}^2[p_1q_1 + p_2q_2]}{d^2} = n = \frac{[1.96]^2[0.2 * 0.8 + 0.25 * 0.75]}{[0.05]^2}; n \geq 533.9824$$

The researcher should draw a sample size of at least 534 from each population.

Exercise 12

1. A sample of 200 measurements of breaking strength of cotton threads gave a mean of 10 ounces and a SD of 1.5 ounces. Find 95% and 99% confidence limits for the mean breaking strength. Write your comment based on the 2 interval estimations.
2. The weight of cement packed bags is distributed normally with a SD of 0.02 kg. A sample of 25 bags is packed up at random, and the mean weight of cement in these 25 bags is only 49.7 kg. Find the confidence interval for the mean weight of cement in filled bags by assuring 10% level of significance.
3. Sodium Vapour Lamps were tested to estimate the life of such a lamp. The life of these 100 lamps exhibited a mean of 10,000 hours with a SD of 500 hours. Construct a 95% confidence interval for the true mean life of a sodium vapour lamp.
4. The daily wages of a random sample of farm labourers are 14, 17, 14.5, 22, 27, 16.5, 19.5, 21, 18, and 22.5. (a) What is the best estimate of the mean daily wages of all farm labourers? (b) What is the standard error of the mean? (c) Compute the interval estimate by considering 99% confidence level.
5. A large retailer issued maintenance contracts for the refrigerators he sells. A random sample is issued to determine the frequency of service calls requiring new parts. A random sample of 100 calls is drawn from a total of 1250 calls recorded in 1 month, out of the 100 sampled calls 42 required new parts. Compute 90% confidence interval estimate of the universe percent.
6. In a large consignment of apples, a random sample of 500 apples revealed that 65 apples were bad. Prove that 99.73% of bad apples in the consignment certainly lies in the interval (0.085, 0.175).
7. An inspector wants to estimate the weight of detergent in packets filled by an automatic filling machine. He wants to be 95% confident that his estimate is not away from the true mean weight of detergent by more than 10 g. What should be the minimum sample size be if it is known that the SD of the weight of detergent filled by that machine is 100 g?
8. A frozen food company wishes to know the mean length of ears of corn received in a large shipment. A random sample of 200 is collected and the ears measured. The arithmetic mean of the lengths is found to be 8.8 inches and the population has an SD of 1.5 inches. What are the 95% confidence limits for μ?
9. The quality-control supervisor of a large manufacturing firm wishes to estimate the mean weight of 5500 packages of raw material. A simple random sample of 250 packages yields a mean of 65 pounds. The population SD is 15 pounds. Construct a confidence interval for the unknown population mean μ. Assume that a 95% confidence interval is satisfactory.
10. Assume that you are the quality-control supervisor for a wire manufacturing company. Periodically you select a sample of wire to test for breaking strength. Experience has shown that the breaking strengths of certain type of wire are normally distributed with an SD of 200 pounds.
11. An advertising firm wants to estimate the average amount of money a certain type of store spent advertising during the past year. Experience has shown the

population variance to be about 1,800,000. How large a sample should the advertising firm take for the estimate to be within \$500/– of the actual mean with 95% confidence?

12. A psychologist wants to construct an interval estimate of the mean of a certain population of employees. The estimate is to be within 5 points of the true mean with 95% confidence. Previous experience indicates that the IQs for the population of interest are approximately normally distributed with a variance of 100. The psychologist wants to know how large a sample to draw from the population.
13. For 2 populations of drivers, an insurance executive wants to estimate the difference in the proportion who regularly wear seat belts. A confidence coefficient of 0.95 and an interval width of 0.12 are desired. Estimates of P_1 and P_2 are 0.25 and 0.18, respectively. How large should the samples be ($n_1 = n_2$)?
14. For 2 populations of employees, an industrial psychologist wishes to estimate the difference in the population proportions who have been sexually harassed at their place of employment. A confidence coefficient of 0.90 and an interval width of 0.14 are desired. How many employees should be selected from each population ($n_1 = n_2$)?
15. A market research firm wants to estimate the proportion of households in a certain area that has colour TV sets. The firm would like to estimate P is within 0.05 with 95% confidence. No estimate of P is available.
16. A sample 100 calculators are taken from a shipment of general electronics calculators, of which 55% are in working condition. Find the 99% confidence interval for the proportion of nondefective calculators produced by general electronics.
17. A manufacturer produces a synthetic fibre at 2 factories located in different parts of the country. Every effort is made to maintain uniformity of production between the 2 factories with respect to the mean-breaking strength of the fibre. To determine whether the 2 factories are maintaining uniformity of production, the manufacturer selects a sample of 25 specimens from the factory 1 and a sample of 16 specimens from factory 2. The mean-breaking strength of the sample from the factory 1 is 22 pounds and of factory 2 is 20 pounds. The variance in both factories is known to be 10 pounds. The populations are normally distributed. Construct the confidence interval for the population mean with a 5% level of significance.
18. An agency conducts a survey to study the characteristics of the subscribers to 2 newspapers. A random sample of 500 subscribers to newspaper A reveals that 300 have annual incomes in excess of \$50,000. In the case of newspaper B, 200 out of a random sample of 500 subscribers have annual incomes in excess of \$50,000. Construct a 95 % confidence interval for the difference between the 2 proportions of subscribers with annual incomes in excess of \$50,000.
19. Doctors who have developed a new drug for the treatment of a certain disease treat a group of 400 patients suffering from the disease with it. They treat another group of 400 patients with an alternative drug. At the end of 2 weeks, 320 of the patients receiving the new drug recover, whereas 240 of those taking the alternative drug recover. Construct the 95% confidence interval for the difference between the true proportions of patients who might be expected to respond to the 2 drugs.

13

Hypothesis Testing, Parametric Tests, Distribution Tests, and Tests of Significance

13.1 Introduction

To help decision makers decide about a population, the data contained in a sample from that population is examined.

To make a decision regarding the population parameter based on the sample information, we are supposed to make an assumption about the population parameters. The assumption made about the population is referred to as a 'hypothesis'. This assumption may be true or false. The methodology that helps to conclude whether the assumption made is true is called 'hypothesis testing'. It can be classified into a null hypothesis (H_0) and an alternative hypothesis (H_1).

13.2 Null Hypothesis (H_0)

According to R. A. Fisher a null hypothesis can be defined as 'the hypothesis that is tested for possible rejection under the assumption that it is true'.

In other words, H_0 asserts that there is no significant difference between the value of the population parameter being tested with the value of the statistic evaluated from a sample drawn.

The null hypothesis normally specifies one of the parameters of the population of interest; the term 'null hypothesis' reflects the idea that this is a hypothesis of no difference. Hence, H_0 always includes a statement of equality.

13.3 Alternative Hypothesis (H_1)

H_1 refers to the alternative available when the null hypothesis must be rejected. Let us assume a situation in which you need to test a hypothesis about a population. If you want to decide whether your sample data provide sufficient evidence to indicate that the population mean is not equal to the value μ_0, your null hypothesis is

Case 1: H_0: $\mu = \mu_0$ and H_1: $\mu \neq \mu_0$

(The alternate hypothesis refer the complement of null hypothesis.)

Here, H_1 is known as a two-sided or (two-tailed) alternative (refer to Figure 13.1).

Case 2: Suppose you raise the question, 'do the sample data provide sufficient evidence to indicate that the population mean is greater than μ_0?'

$$H_1: \mu > \mu_0; H_0 : \mu \leq \mu_0$$

Here, H_1, is known as one-sided (one-tailed) or right-tailed alternative (refer to Figure 13.2).

Case 3: Suppose you raise the question, 'do the sample data provided is sufficient evidence to indicate that the population mean is less than μ_0?'

$$H_1: \mu < \mu_0; H_0: \mu \geq \mu_0$$

Here, H_1 is known as one-sided (one-tailed or left-tailed) alternative (refer to Figure 13.3).

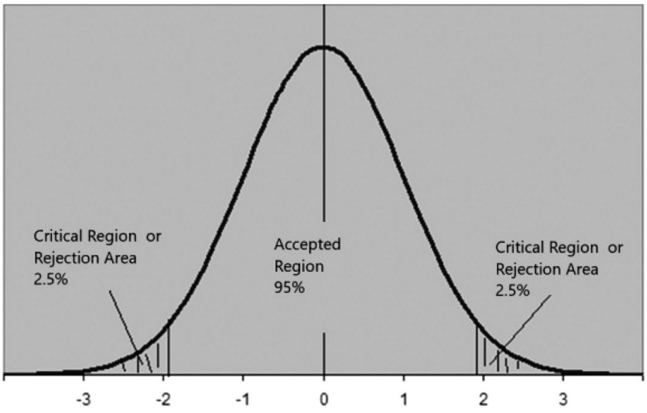


FIGURE 13.1
Two-tailed test.

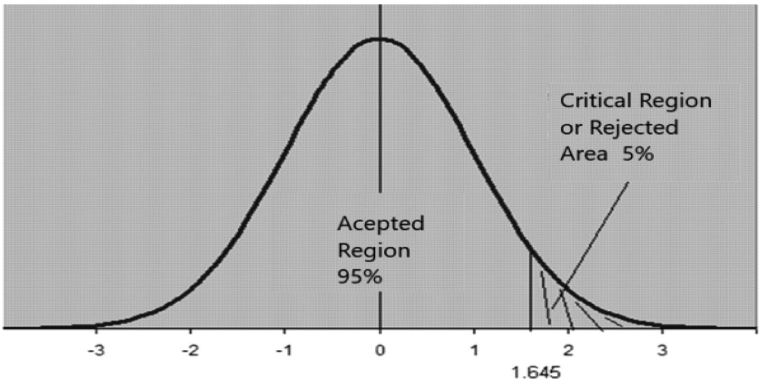


FIGURE 13.2
One-tailed test [right-tailed].

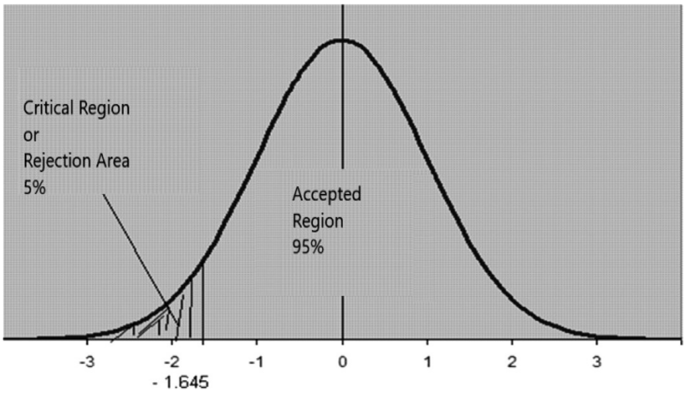


FIGURE 13.3
One-tailed test [left-tailed].

NOTE: To avoid the status of confusion and to decide the alternatives easily, first decide H_1, and then decide H_0. (H_0 is the complement of H_1.)

Hypothesis tests are either one-tailed or two-tailed. This is normally decided by the nature of H_1.

If H_1: $\mu > \mu_0$ or H_1: $\mu < \mu_{0,}$ the test is one-tailed (right-tailed or left-tailed).

On the other hand, if H_1: $\mu \neq \mu_0$, the test is both right- and left-tailed and hence, two-tailed.

13.4 Type I and Type II Errors

Type I Error

Rejection of H_0 where it is true; where the probability of a type I error (given that H_0 is true) is denoted by α, that is P (Reject H_0/H_0 True) = α.

Type II Error

Acceptance of H_0 when it is false; where the probability of a type II error (given that H_1 is true) is denoted by β, that is, P (Accept H_0/H_1 True) = β.

The same can be referred to as

Evaluated Actual	H_0 Accepted	H_0 Rejected
H_0 True	No error	Type I error
H_0 False	Type II error	No error

13.5 Meaning of Parametric and Nonparametric Test

13.5.1 Parametric Test

The parametric statistical test is a test whose model species certain conditions about the parameters of the population from which the sample is drawn. Sample statistics will be used to test the hypothesis that will be made about certain universe parameters. The nature of population distribution from which the sample is drawn is known. Few of the parametric tests are *Z*-test, *t*-test, etc.

13.5.2 Nonparametric Test

Nonparametric tests are often referred to as *distribution free* test because they do not rely on assumptions that the data are drawn from a given probability distribution. The term 'nonparametric statistic' can also refer to a statistic. Nonparametric methods are widely used for studying populations that take on a ranked order (such as movie reviews receiving one to four stars). The use of nonparametric methods may be necessary when data have a ranking but no clear numerical interpretation, such as when assessing preferences. Because nonparametric methods make fewer assumptions, their applicability is much wider than that of the corresponding parametric methods. Chi-squared test falls into this category.

13.6 Selection of Appropriate Test Statistic

Population Distribution	Population Variance, σ^2	Sample Size, n	Appropriate Test Statistic
Follows normal	Known	Any size	*Z*-test
Follows normal	Unknown	<30	*t*-test
Follows normal	Unknown	≥30	*t*-test or *Z*-test
Any	Known or unknown	≥30	*Z*-test

Almost all the tests in this section are based on normal distribution theory, assuming either that the samples come from normal population, or n is sufficiently large to justify normal approximation.

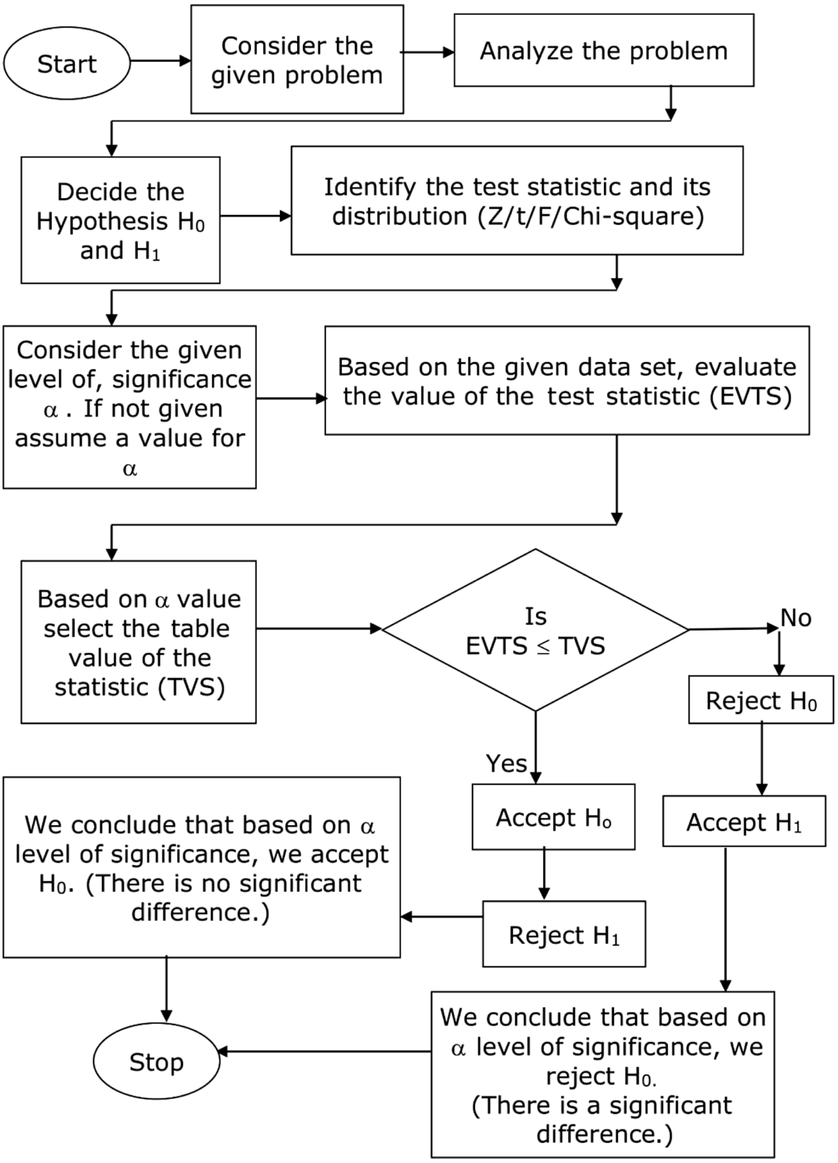
To study the difference between two population variances, use an *F*-test or *Z*-test.

To study the difference of three or more population means, one can make use of analysis of variance (ANOVA) test.

For goodness of fit, independence of attributes, and test for specified population variance, use of the chi-squared test (nonparametric test).

13.7 Methodology of Statistical Testing

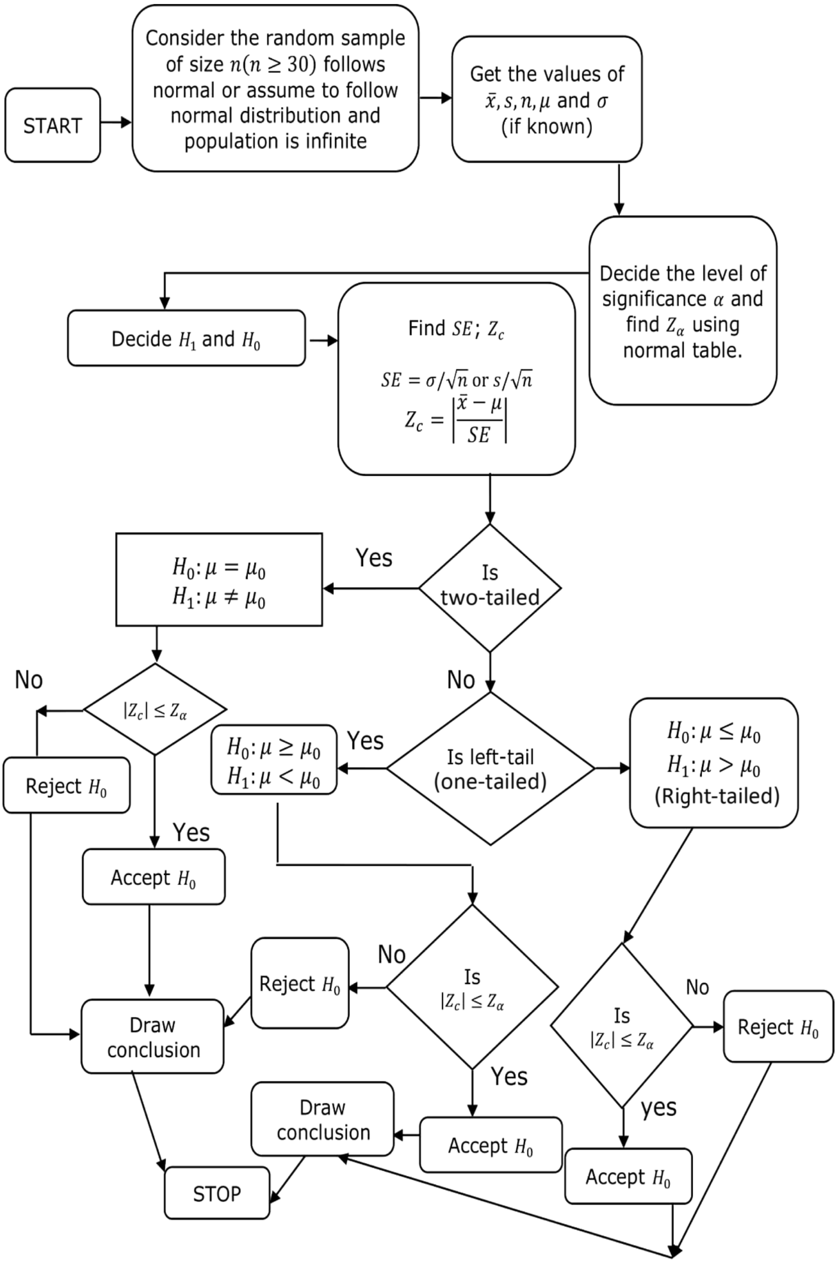
The sequential step to be followed in the case of hypothesis testing is depicted clearly in Flowchart 13.1.


Start
Consider the given problem
Analyze the problem
Decide the Hypothesis H0 and H1
Identify the test statistic and its distribution (Z/t/F/Chi-square)
Consider the given level of, significance α . If not given assume a value for α
Based on the given data set, evaluate the value of the test statistic (EVTS)
Based on α value select the table value of the statistic (TVS)
Is EVTS ≤ TVS
No
Reject H0
Yes
Accept H0
Accept H1
Reject H1
We conclude that based on α level of significance, we accept H0. (There is no significant difference.)
We conclude that based on α level of significance, we reject H0. (There is a significant difference.)
Stop

FLOWCHART 13.1
Hypothesis testing.

13.8 Test for a Specified Mean: Large Sample

The sample is said to be larger if the size $n \geq 30$. The sampling distribution corresponding to large samples is approximated to normal distribution. The decision rule can be formulated based on the following tabular value of the Z statistic with the level of significance α. Refer to Flowchart 13.2.



FLOWCHART 13.2
Test for a specified mean (large sample).

	Critical Values for		
Level of Significance α	**Two-Tailed**	**Right-Tailed**	**Left-Tailed**
1%	$\lvert z\alpha \rvert > 2.58$	$z\alpha > 2.33$	$z\alpha < -2.33$
5%	$\lvert z\alpha \rvert > 1.96$	$z\alpha > 1.645$	$z\alpha < -1.645$
10%	$\lvert z\alpha \rvert > 1.645$	$z\alpha > 1.28$	$z\alpha < -1.28$

Note: Without any reference to the level of significance, the null hypothesis can be rejected when $\lvert z \rvert > 3$.

Example:

An automatic machine was designed to pack exactly 2.0 kg of oil. A sample of 100 tins was examined to test the machine. The mean weight was found to be 1.94 kg with a SD of 0.1 kg. Is the machine working properly? (Use 5% level of significance.)

Step 1: Given the values:

Population	**Sample**
Mean = μ = 2 kg	Mean = $\bar{x}$ = 1.94 kg s = 0.1 kg n = 100

Step 2: Framing the hypothesis

$$H_0: \mu = 2 \text{ kg}; H_1: \mu \neq 2 \text{ kg}$$

Step 3: Defining the test statistic

Since the parameter of interest is population mean, μ, the relevant statistic is to be evaluated from sample mean, $\bar{x}$. When the sampled population is normally distributed, the sampling distribution $\bar{x}$ is also normal with mean μ and SD = $\frac{\sigma}{\sqrt{n}}$.

The test statistic to be evaluated is Z_c [Z – calculated value] and it is defined as,

$$Z_c = \frac{\bar{x} - \mu}{SE[\bar{x}]}$$

where $SE(\bar{x}) = \frac{s}{\sqrt{n}}$ since the population SD is unknown.

Step 4: Defining the significance level

Here, decision sets the level of significance at $\alpha = 0.05$.

Since H_1: $\mu \neq 2$ implies that the test is a two-tailed one.

According to the Z-table, the critical value of Z_t is $Z_t(\alpha) = Z_t(0.05) = 1.96$ (two-tailed).

We see that if Z_c of a sample statistic lies between –1.96 and 1.96, then we are 95% confident that H_0 is true. ($P(-1.96 \leq z \leq 1.96) = 0.95$)

(Refer to Figure 13.1.)

Step 5: Evaluate $SE(\bar{X}) = \frac{s}{\sqrt{n}} = \frac{0.1}{\sqrt{100}} = \frac{0.1}{10} = 0.01$

$$Z_c = \frac{\bar{x} - \mu}{SE(\bar{x})} = \frac{1.94 - 2}{0.01} = -6; \left|z_c\right| = \left|-6\right| = 6$$

Step 6: Statistical decisions

Since $\left|z_c\right| = 6$, lies in the critical region $|z| > 1.96$, according to the decision rule, we reject H_0.

We can say that we reject the null hypothesis, H_0, because 6 is greater than 1.96.

Step 7: Conclusion

We conclude that at 5% of level of significance, the mean packaging weight of the population oil packing cannot be taken as 2 kg. The machine is not functioning properly.

Example:

The mean breaking strength of the cables supplied by a manufacturer is 1800 with an SD 100. By a new technique in the manufacturing process, it is claimed that the breaking strength of the cables has increased. To test this claim, a sample of 50 cables is tested. It is found that the mean breaking strength is 1850. Can we support the claim based on 1% level of significance?

Step 1: Given the values:

Population	**Sample**
$\mu = 1800$	Mean = $\bar{x} = 1850$
$SD = \sigma = 100$	Size = n = 50

Step 2: Framing the hypothesis

$$H_0\text{: } \mu = 1800;\ H_1\text{: } \mu \succ 1800 \text{ (Because 1850 > 1800)}$$

Step 3: Defining the test statistic

Because the parameter of interest is population mean, μ, the relevant statistic is to be evaluated from sample mean, $\bar{x}$. When the sample population is normally distributed, the sample distribution $\bar{x}$ is also normal with mean μ and SD = $\frac{\sigma}{\sqrt{n}}$.

The test statistic to be evaluated is Z_c, and the same is defined as

$$Z_c = \frac{\bar{x} - \mu}{SE[\bar{x}]}; \text{ where } SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Step 4: Defining the significance level

Here, decision sets the level of significance at $\alpha = 0.01$.

H_1: $\mu > 1800$, implies that the test is a one-tailed (right-tailed) test (refer to Figure 13.4). According to the table, the critical value of Z_t is $z_t(\alpha) = z_t(0.01) = 2.33$.

We see that if Z_c of the sample statistic is less than or equal to 2.33, then we are 99% confident that H_0 is true ($P(Z \leq 2.33) = 0.99$).

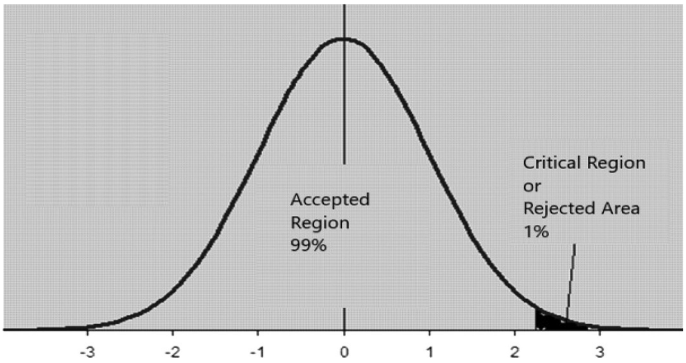


FIGURE 13.4
Right-tailed test with $\alpha = 0.01$.

Step 5: Evaluate $SE(\bar{x})$ and Z_c

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{50}} = 14.1421$$

$$z_c = \frac{\bar{x}-\mu}{SE(\bar{x})} = \frac{1850-1800}{14.1421} = 3.5355$$

Step 6: Statistical decisions

Since $|Z_c| = 3.5355$, which lies in the critical region $Z > 2.33$, according to the decision rule, we reject H_0.

Step 7: Conclusion

We conclude that at 1% of the level of significance, the mean breaking strength of cables has increased. The claim is justified.

Example:

An insurance agent has claimed that the average age of policyholder who insure through him is less than the average for all agents, which is 30.5 years. A random sample of 100 policyholders who had insurance through him gave the following age distribution.

Age	Number of Persons
16–20	12
21–25	22
26–30	20
31–35	30
36–40	16
Total	100

Test the agent's claim with a 5% level of significance.

Step 1: Evaluate the sample mean and SD based on the sample data with the usual procedure. Because the class integral is not continuous, convert it.

The cross difference = 1. Half of the cross difference = ½.
Modify the class interval as (L – ½, U + ½). The modified problem is

Age	Frequency, [f]	Mid Value [X]	$h = 5, A = 28,$ $d = \frac{x-A}{h}$	d^2	fd	fd^2
15.5–20.5	12	18	–2	4	–24	48
20.5–25.5	22	23	–1	1	–22	22
25.5–30.5	20	28	0	0	0	0
30.5–35.5	30	33	1	1	30	30
35.5–40.5	16	38	2	4	32	64
Total	100				16	164

$$\bar{x} = A + h\left\{\frac{\sum_{i=1}^{5} f_i d_i}{\sum_{i=1}^{5} f_i}\right\}$$

By definition

$$\bar{x} = 28 + 5\left\{\frac{16}{100}\right\}$$

$$\bar{x} = 28.8 \text{ year}$$

$$\text{SD} = s = 5 \times \sqrt{\frac{\sum_{i=1}^{5} f_i d_i^2}{n} + \left(\frac{\sum_{i=1}^{5} f_i d_i}{n}\right)^2}$$

$$s = 5 \times \sqrt{\frac{164}{100} + \left(\frac{16}{100}\right)^2} = 6.45 \text{ years}$$

Now, we have the sample statistic,
Mean = $\bar{x}$ = 28.8 years; SD = s = 6.45 years

$$n = 100$$

Let μ be the population mean.
Step 2: Framing the hypothesis

$$\text{H}_1\text{: } \mu < 30.5 \text{ years; H}_0\text{: } \mu \geq 30.5 \text{ years}$$

Step 3: Defining the test statistic

Because the sample is a large one, the corresponding test statistic is Z. It can be defined as,

$$Z_c = \frac{\bar{x} - \mu}{SE[\bar{x}]}; \text{ where } SE[\bar{x}] = \frac{s}{\sqrt{n}}.$$

Step 4: Defining the significance level

The level of significance α is given as 0.05. Since H_1: $\mu < 30.5$ implies that the test is a one-tailed (left-tailed) one (refer to Figure 13.3), according to the table, the critical value of $Z_t(\alpha) = Z_t(0.05) = -1.645$.

We see that if Z_c of the sample statistic $Z_c \geq -1.645$, then we are 95% confident that H_0 is true.

$$[P(Z_c \geq -1.645) = 0.95]$$

Step 5: Evaluate $SE(\bar{x})$ and $Z_{c.}$

$$SE[\bar{x}] = \frac{s}{\sqrt{n}} = 6.45/\sqrt{100} = 0.645; Z_c = \frac{\bar{x} - \mu}{SE[\bar{x}]} = \frac{28.8 - 30.5}{0.635} = -2.636$$

Step 6: Statistical decisions

Since the value of $Z_c = -2.636$, lies in the critical region, according to the decision rule, we reject H_0.

Step 7: Conclusion

At 5% level of significance, we conclude that the insurance agents claim is valid.

Example:

The quality control department of food-processing firm specifies that the mean net weight per package of cereal should not be less than 20 ounces. Experience has known that the weights are approximately normally distributed with SD of 1.5 ounces. A random sample of 15 packages yields a mean weight of 19.5 ounces. Is this sufficient evidence to indicate that the true mean weight of the package has decreased?

Step 1: Given the values:

Population	Sample
$\sigma = 1.5$ ounces	$n = 15$
	$\bar{x} = 19.5$ ounces

Step 2: Framing the hypothesis

$$H_0: \mu \geq 20 \text{ ounces}; H_1: \mu < 20 \text{ ounces}$$

Step 3: Defining the test statistic

Because the population is approximately normally distributed, and we know the population SD – σ, the test statistic to be evaluated is Z_c and it is defined as,

$$z_c = \frac{\overline{x} - \mu}{SE(\overline{x})}; \text{ where } SE(\overline{x}) = \frac{\sigma}{\sqrt{n}}\cdot$$

Step 4: Defining the significance level

Because the level of significance α is not stated, let us assume that $\alpha = 0.05$.

Since $H_1 \prec 20$ implies that the test is a one-tailed (left-tailed) test (refer to Figure 13.3), according to the table, the critical value of $Z_t(\alpha) = Z_{t(0.05)} = -1.645$.

We see that if Z_c of the sample statistic is greater than or equal to –1.645 (because the value of $Z_{0.05}$ is negative), then we are 95% confident that H_0 is true ($P(-1.645 \leq Z) = 0.95$).

Step 5: Evaluate $SE(\overline{x})$ and Z_c

$$SE(\overline{x}) = \frac{\sigma}{\sqrt{n}} = \frac{1.5}{\sqrt{15}} = 0.3873; z_c = \frac{\overline{x} - \mu}{SE(\overline{X})} = \frac{19.5 - 20}{0.3873} = -1.291; Z_c = -1.291$$

Step 6: Decision rule

Since the value of $Z_c = -1.291$ lies in the accepted region, according to the decision rule, we accept H_0.

Step 7: Conclusion

We conclude that at the 95% level of significance the true mean weight has decreased.

Example:

The weight of cement bags produced in Beta Cement Company follows a normal distribution whose population is finite and its size is 1000. The expected mean of the weight of the cement bags for sales of this population is 65 kg and its variance is unknown. The sales manager of the firm claims that the mean weight of the cement bags is significantly more than the expected weight of the population. So, the purchase manager of Alpha Construction Company, who places the order for cement bags with the Beta Cement Company, has selected a random sample of 64 bags and its mean and variance are found to be 62 and 2.25 kg, respectively. Verify the intuition of the sales manager of the cement company at a significance level of 0.05.

Step 1: Given the values:

Population	**Sample**
$N = 1000$	$n = 64$
$\mu = 65$ kg	$\overline{x} = 62$ kg
	$s = 2.25$ kg

Step 2: Framing the hypothesis

$$H_0: \mu = 65 \text{ kg}; H_1: \mu > 65 \text{ kg}$$

Step 3: Defining the test statistic

Because the population size is large and is assumed that it follows normal, and we know the sample, SD – s, the test statistic to be evaluated is Z_c, and it is defined as,

$$Z_c = \frac{\bar{x}-\mu}{SE(\bar{x})}; \text{ where } SE(\bar{x}) = \frac{s}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}.$$

(Since σ is not known and the sample is finite sample.)

Step 4: Defining the significance level

Here, the decision sets the level of significance $\alpha = 0.05$.

Since $H_1 > 65$ implies that the test is one-tailed (right-tailed) test (refer to Figure 13.2), according to the table, the critical value of $z_t(\alpha) = z_t(0.05) = 1.645$.

We see that if Z_c of the sample statistic is less than or equal to 1.645, then we are 95% confident that H_0 is true $(P(Z \le 1.645) = 0.95)$.

Step 5: Evaluate $SE(\bar{X})$ and Z_c

$$SE(\bar{X}) = \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{2.25}{\sqrt{64}}\sqrt{\frac{1000-64}{1000-1}} = \frac{2.25}{\sqrt{64}}\sqrt{\frac{936}{999}} = \frac{2.25}{8}\left[0.968\right]$$

$$SE(\bar{X}) = 0.27225$$

$$Z_c = \frac{\bar{x}-\mu}{SE(\bar{x})} = \frac{62-65}{0.27225} = -11.0192; |z_c| = |-11.0192| = 11.0192$$

Step 6: Statistical decisions

Since $|z_c| = 11.0192$ lies in the critical region $Z_c > 1.645$, according to the decision rule, we reject H_0.

Step 7: Conclusion

We conclude that at the 5% level of significance the mean weight of the cement bags is more than the expected weight of the population.

Example:

A market research firm is interested in the amount that households in a certain town spent on groceries each week. The firm believes that the average amount spent per household each week is less than $90. A random sample of 100 households yields a mean of $88 and a SD of $10. Do these data support the firm's belief?

Step 1: Given the values:

Population	**Sample**
$\mu = 90$	$n = 100$
	$\bar{x} = 88$
	$s = 10$

Step 2: Framing the hypothesis

$$H_0\text{: } \mu \ge 90; H_1\text{: } \mu < 90$$

Step 3: Defining the test statistic

Since $n = 100 \geq 30$, it refers to a large sample. We assume that this follows a normal distribution. Hence, the statistic to be evaluated is Z_c, and the same is defined as

$$Z_c = \frac{\bar{x} - \mu}{SE(\bar{x})}; \text{ where } SE(\bar{x}) = \frac{s}{\sqrt{n}}; \text{ since } \sigma \text{ is unknown.}$$

Step 4: Defining the significance level

Because the level of significance α is not stated, let us assume that $\alpha = 0.05$.

Since $H_1 < 90$, implies that the test is a one-tailed (left-tailed) test (refer to Figure 13.3), according to the table, the critical value $Z_t[\alpha]$ is $Z_t[\alpha] = Z_t\,[0.05] = -1.645$.

We see that if Z_c of the sample statistic is greater than or equal to −1.645 (because the value of $Z_t[0.05]$ is negative), then we are 95% confident that H_0 is true ($P(-1.645 \leq Z) = 0.95$).

Step 5: Evaluate $SE(\bar{X})$ and Z_c

$$SE(\bar{X}) = \frac{s}{\sqrt{n}} = \frac{s}{\sqrt{100}} = \frac{10}{10} = 1$$

$$Z_c = \frac{\bar{x} - \mu}{SE(\bar{X})} = \frac{88 - 90}{1} = -2.$$

Step 6: Statistical decisions

Since the value of $Z_c = -2$ lies in the critical region, according to the decision rules, we reject H_0.

Step 7: Conclusion

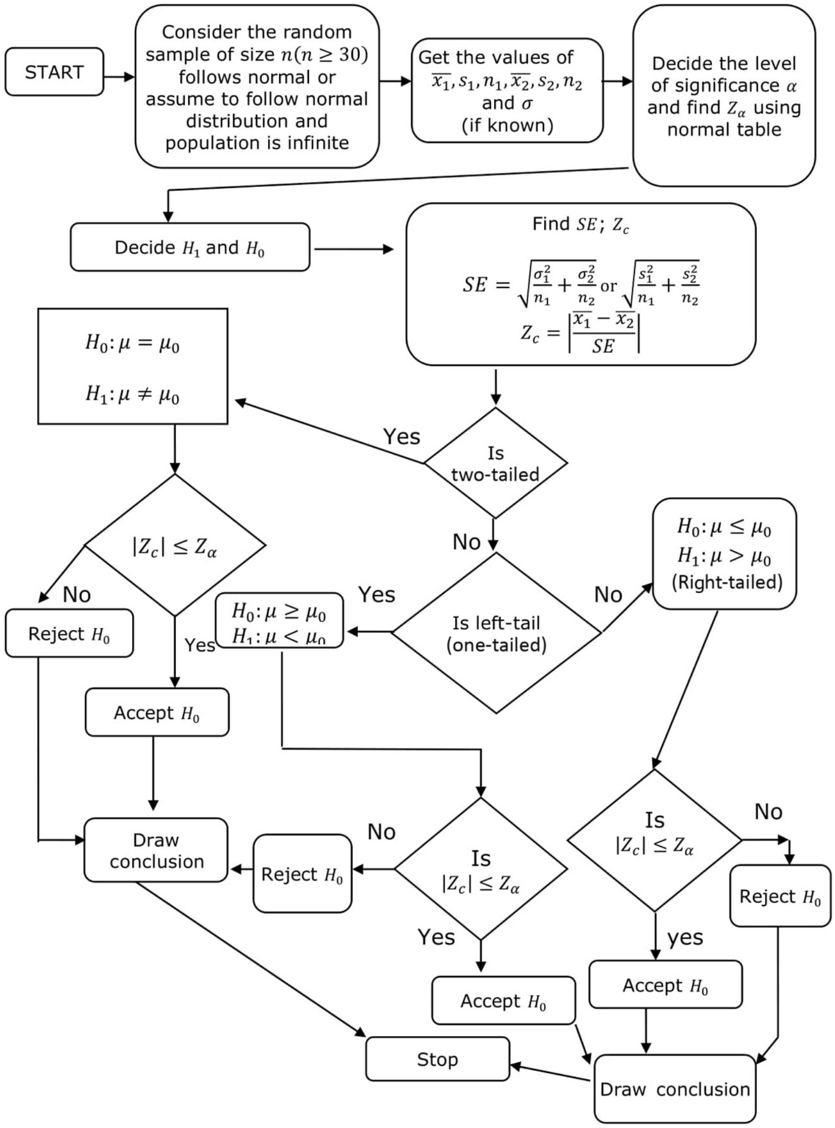
We conclude that at the 5% level of significance the average amount spent per household each week is less than $90.

13.9 Test for Equality of Two Populations: Large Sample

Refer to Flowchart 13.3.

Example:
As part of an investigation of employee turnover, an industry-wide survey of people in sales management positions gave the number of years of experience in sales-related positions. The data is as follows:

	Males	Females
Sample size	80	70
Mean (years)	21.7	18.5
SD (years)	9.3	4.8


START
Consider the random sample of size $n(n \geq 30)$ follows normal or assume to follow normal distribution and population is infinite
Get the values of $\overline{x_1}, s_1, n_1, \overline{x_2}, s_2, n_2$ and σ (if known)
Decide the level of significance α and find Z_α using normal table
Decide H_1 and H_0
Find SE; Z_c
$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ or $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$Z_c = \left|\frac{\overline{x_1} - \overline{x_2}}{SE}\right|$
$H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$
Yes
Is two-tailed
No
$|Z_c| \leq Z_\alpha$
No
Reject H_0
Yes
Accept H_0
$H_0: \mu \geq \mu_0$
$H_1: \mu < \mu_0$
Yes
Is left-tail (one-tailed)
No
$H_0: \mu \leq \mu_0$
$H_1: \mu > \mu_0$
(Right-tailed)
Draw conclusion
Reject H_0
No
Is $|Z_c| \leq Z_\alpha$
Yes
Accept H_0
Is $|Z_c| \leq Z_\alpha$
No
Reject H_0
yes
Accept H_0
Stop
Draw conclusion

FLOWCHART 13.3
Test for difference of two population means (large sample).

Test with 5% level of significance, the difference of means.

Step 1: Given the values:

Sample 1	Sample 2
Mean = $\bar{x}_1 = 21.7$	Mean = $\bar{x}_2 = 18.5$
SD = $s_1 = 9.3$	SD = $s_2 = 4.8$
Size = $n_1 = 80$	Size = $n_2 = 70$

Let μ_1, and μ_2 stand for the means of population 1 and population 2, respectively.

Step 2: Framing the hypothesis

$$H_0\colon \mu_1 = \mu_2;\ H_1\colon \mu_1 \neq \mu_2$$

Step 3: Defining the test statistic

Since $n_1 = 80$ and $n_2 = 70$ and both are more than 30, the samples are categorized as large samples. We assume that this follows a normal distribution. Hence, the test statistic to be evaluated is Z_c, and it can be defined as,

$$Z_c = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)};\ \text{where } SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1{}^2}{n_1} + \frac{s_2{}^2}{n_2}};\ \text{since } \sigma_1 \text{ and } \sigma_2 \text{ are unknown.}$$

NOTE: Whenever H_0: $[\mu_1 - \mu_2] \neq 0$, we use the modified formula for the evaluation of Z_c.

$$Z_c = \frac{[\bar{x}_1 - \bar{x}_2] - [\mu_1 - \mu_2]}{SE(\bar{x}_1 - \bar{x}_2)}.$$

Step 4: Defining the significance level

Here, the decision sets the level of significance at $\alpha = 0.05$. Since $H_1 : \mu_1 \neq \mu_2$, it implies that the test is a two-tailed one (refer Figure 13.1). According to the table, the critical value of $Z_t(\alpha) = Z_t(0.05) = 1.96$.

We see that if Z_c of a sample statistic lies between –1.96 and +1.96, then we are 95% confident that H_0 is true ($P(-1.96 \le Z \le 1.96) = 0.95$).

Step 5: Evaluate $SE(\bar{x})$ and Z_c

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1{}^2}{n_1} + \frac{s_2{}^2}{n_2}} = \sqrt{\frac{9.3^2}{80} + \frac{4.8^2}{70}} = \sqrt{1.4102}$$

$$SE(\bar{x}_1 - \bar{x}_2) = 1.1876$$

$$Z_c = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x})} = \frac{21.7 - 18.5}{1.1876}$$

$$Z_c = 2.6946$$

Step 6: Statistical decisions

Since $Z_c = 2.6946$ lies in the critical region, $|z| > 1.96$, according to the decision rule, we reject H_0.

Step 7: Conclusion

At the 5% level of significance, we conclude that there is a significant difference between the two populations mean.

Example:

A random sample of 1000 mill workers at Madurai showed their mean wages to be \$47 per month with a SD of \$28. A sample of 1500 mill workers in Trichy showed their mean wages to be \$49 per month with a SD of \$40. On the basis of these data would you say that the mean wage of mill workers in Trichy is higher than that of those in Madurai?

Step 1: Given the values:

Sample 1	**Sample 2**
Size = $n_1 = 1500$	Size = $n_2 = 1000$
Mean = $\bar{x}_1$ = \$49	Mean = $\bar{x}_2$ = \$47
SD = s_1 = \$40	SD = s_2 = \$28

Let μ_1 and μ_2 stand for the means of population 1 and population 2, respectively.

Step 2: Framing the hypothesis

$$H_1: \mu_1 > \mu_2; H_0: \mu_1 \leq \mu_2$$

Step 3: Defining the test statistic

Since $n_1 = 1500$ and $n_2 = 1000$ and both more than 30, it implies that the sample is large. We assume that this follows a normal distribution. Hence, the test statistic to the evaluated is Z_c, and the same can be defined as,

$$Z_c = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)}; \text{ where } SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}; \text{ since } \sigma_1 \text{ and } \sigma_2 \text{ are unknown.}$$

Step 4: Defining the significance level

Because the level of significance is not given, let us assume that $\alpha = 0.05$.

Since $H_1 : \mu_1 > \mu_2$, it implies that the test is a right-tailed one (refer Figure 13.2). According to the table, the critical value of $Z_t(\alpha) = Z_t(0.05) = 1.645$.

We see that if Z_c of the sample statistic is less than or equal to 1.645, then we are confident that H_0 is true ($P(z \leq 1.645) = 0.95$).

Step 5: Evaluate $SE(\bar{x})$ and Z_c

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}} = \sqrt{\frac{40^2}{1500} + \frac{28^2}{1000}} = \sqrt{1.8507} = 1.3604$$

$$Z_c = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)} = \frac{49 - 47}{1.3604} = 1.4702$$

Step 6: Statistical decisions

Since $Z_c = 1.4702$ lies in the acceptance region, according to the decision rule, we accept H_0.

Step 7: Conclusion

At the 5% level of significance, we conclude that the mean wage of mill workers in Trichy is not higher than that of those in Madurai.

Example:

Rampling Management is infuriated at Corporal Motors' claim that its Americana is superior to Rambling's Futura economy vehicle. Corporal claims that its Americana outlasts Rambling's Futura by at least 10,000 miles. Rambling says that its Futura is at least as good as Corporal's Americana. A consultant was hired to test the Corporal's claim. Samples, based on prototype testing, were taken, and the average lifetimes were found to be as follows:

Particulars	**Rambling [Futura]**	**Corporal [Americana]**
Size	200	150
Mean (miles)	80,000	92,000
SD (miles)	8,000	12,000

Acting as the consultant, he constructs and conducts a hypothesis test, at the 5% level, to evaluate Corporal's claim.

Step 1: Given the values:

Sample 2	**Sample 1**
Rambling	Corporal
$n_2 = 200$	$n_1 = 150$
$\bar{x}_2 = 80{,}000$	$\bar{x}_1 = 9{,}2000$
$s_2 = 8{,}000$	$s_1 = 12{,}000$

Let $\mu_1 \,\&\, \mu_2$ be the two population means of Corporal and Rambling, respectively.

Step 2: Framing the hypothesis

$$H_1 : \mu_1 - \mu_2 > 10000;\ H_0 : \mu_1 - \mu_2 <= 10000$$

Step 3: Defining the test statistic

Because the sample sizes are large, let us assume that the data follows a normal distribution.

The test statistic to be evaluated is Z_c, and it is defined as

$$Z_c = \frac{[\bar{x}_1 - \bar{x}_2] - [\mu_1 - \mu_2]}{SE(\bar{x}_1 - \bar{x}_2)},\ \text{where } SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^{\ 2}}{n_1} + \frac{s_1^{\ 2}}{n_2}}.$$

Step 4: Defining the significance level

Because the level of significance α is not specified, let us assume $\alpha = 0.05$.

$H_1 : \mu_1 - \mu_2 > 10000$ implies that the test is a one-tailed (right-tailed) test. (Refer Figure 13.2).

According to the Z-table, the critical value of $Z_t(\alpha)$ is $Z_t(\alpha) = Z_t(0.05) = 1.645$.

We see that if Z_c of the sample statistic is less than or equal to 1.645, then we are 95% confident that H_0 is true ($P(Z \leq 1.645) = 0.95$).

Step 5: Evaluate $SE(\bar{x}_1 - \bar{x}_2)$ and Z_c

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1{}^2}{n_1} + \frac{s_2{}^2}{n_2}}; \text{ since } \sigma_1 \,\&\, \sigma_2 \text{ are unknown.}$$

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{8000^2}{200} + \frac{12000^2}{150}} = \sqrt{1280000} = 1131.3709$$

$$Z_c = \frac{[\bar{x}_1 - \bar{x}_2] - [\mu_1 - \mu_2]}{SE(\bar{x}_1 - \bar{x}_2)} = \frac{12000 - 10000}{1131.3709} = 1.7678$$

Step 6: Statistical decisions

Since $Z_c = 1.7678 > 1.645$ lies in the rejection region, according to the decision rule, we reject H_0.

Step 7: Conclusion

We conclude that at 5% level of significance the claim is not correct. That is, the Americana does not outlast the Futura by 10,000 miles.

13.10 Test for Population Proportion: Large Sample

Refer to Flowchart 13.4.

Example:

Tossing an unbiased coin 1000 times resulted in 470 heads. Test the hypothesis that the coin is fair. Use 5% level of significance.

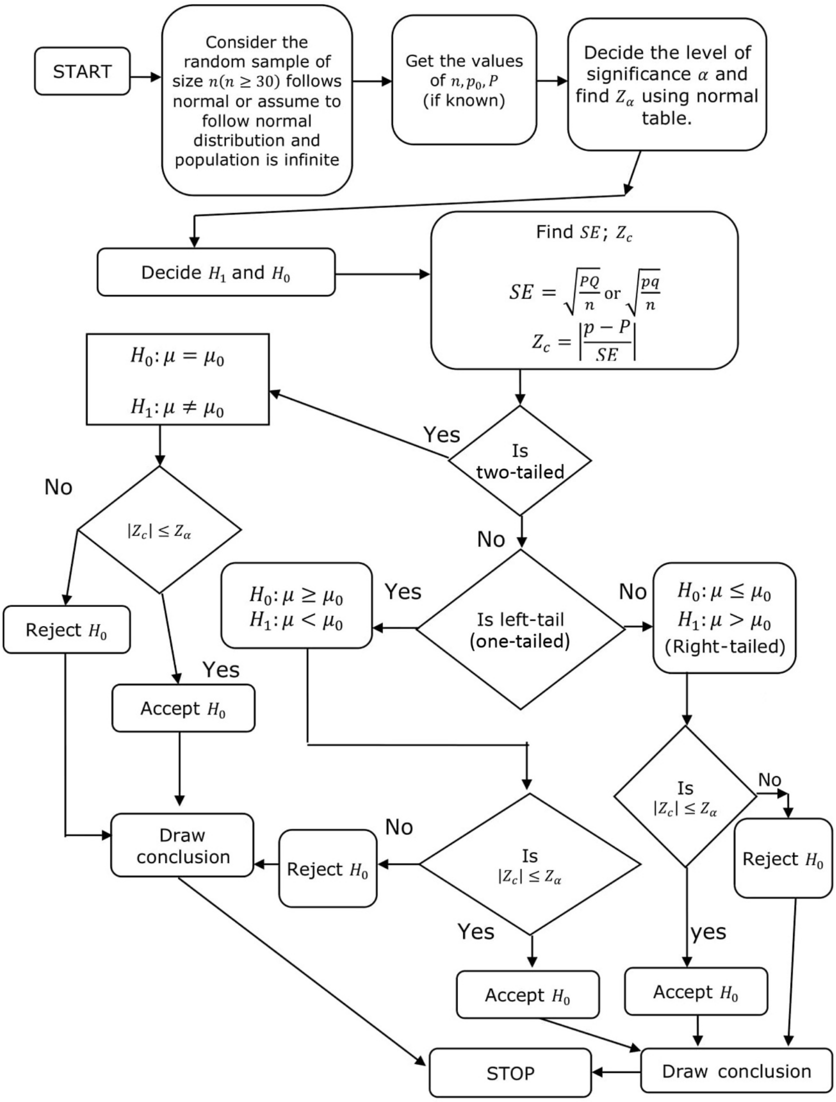
Step 1: Given the values:

Population	Sample
Let P be the population proportion	$p = \frac{470}{1000} = 0.47$ $q = 1 - p = 1 - 0.47 = 0.53$ $n = 1000$

Step 2: Framing the hypothesis

$$H_0 : P = 0.5$$

$$H_1 : P \neq 0.5$$



FLOWCHART 13.4
Test for a specified proportion (large sample).

Step 3: Defining the test statistic

The test statistic to be evaluated is Z_c, and it is defined as

$$Z_c = \frac{p-P}{SE(p)}, \text{ where } SE(p) = \sqrt{\frac{pq}{n}} \; [\because P \text{ is unknown}].$$

Step 4: Significance level

Here, decision sets the level of significance at $\alpha = 0.05$.

Since $H_1 : P \neq 0.5$, it implies that the test is a two-tailed one (refer to Figure 13.1). According to the table, the value of $Z_t(\alpha) = Z_t(0.05) = 1.96$.

We see that of Z_c of a sample statistic lies between –1.96 and +1.96, so we are 95% confident that H_0 is true $(P(-1.96 \le Z \le 1.96)) = 0.95)$.

Step 5: Evaluate the value of $SE(p)$ and Z_c

$$SE(p) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.47 \times 0.53}{1000}} = 0.0158$$

$$Z_c = \frac{p-P}{SE(p)} = \frac{0.47-0.5}{0.0158}$$

$$Z_c = -1.89873; |Z_c| = |-1.89873| = 1.89873$$

Step 6: Statistical decisions

Since $|Z_c| = 1.89873$, lies in the critical region $|z| < 1.96$, according to the decision rule, we accept H_0.

Step 7: Conclusion

At the 5% level of significance, we conclude that the population of the proportion can be 0.5; that is, the coin is fair one.

Example:

In a large sample of 500 items manufactured by the company Bhavana Shree Ltd., the number of defective items was 20. The purchaser claimed that 5% of their items are defective. Is the claim justified? Test it with a 5% level of significance.

Step 1: Given the values:

Population	Sample
$P = 0.05$	$n = 500$
$Q = 1 - P = 1 - 0.05 = 0.95$	$p = 20/500 = 0.04$
	$q = 1 - p = 1 - 0.04 = 0.96$

where $P = P$ (item being defective in the population) = 0.05 and $p = P$ (item being defective in the sample).

Step 2: Framing the hypothesis

$$H_0 : P \ge 0.05$$

$$H_1 : P < 0.05 \text{ [since } 0.04 < 0.05]$$

Step 3: Defining the test statistic

The test statistic to be evaluated is Z_c, and it can be defined as

$$Z_c = \frac{p-P}{SE(p)}, \text{ where } SE(p) = \sqrt{\frac{PQ}{n}}.$$

Step 4: Defining the significance level

Because the level of significance is not given, let us assume that $\alpha = 0.05$.

$H_1 : P < 0.05$ implies that the test is one-tailed (left-tailed) one (refer to Figure 13.3). According to the table, the value of $Z_t(\alpha) = Z_t(0.05) = -1.645$.

We see that if Z_c of a sample statistic satisfies the condition $1.645 \le Z$, then we are 95% confident that H_0 is true $P(-1.645 \le Z) = 0.95$.

Step 5: Evaluate the value of $SE(P)$ and Z_c.

$$SE(p) = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.05 \times 0.95}{500}} = \sqrt{0.000095} = 0.00975$$

$$Z_c = \frac{p-P}{SE(p)} = \frac{0.04-0.05}{0.00975} = -1.0256.$$

Step 6: Statistical decisions

Since $Z_c = -1.20256$ lies in the acceptance region, according to the decision rule we accept $H_{0.}$

Step 7: Conclusion

At the 5% level of significance, we conclude that the claim is justified.

Example:

The president of a certain firm, concerned about the safety record of the firm's employees, sets aside $15,000 a year for safety education. The firm's accountant believes that more than 75% of similar firms spend more than $15,000 a year on safety education. To prove this claim, he collected a simple random sample of size 60 firms. Out of the 60 firms, 50 firms accepted that they spend more known $15,000 per year on safety education. Comment on the claim of the accountant.

Step 1: Given the values:

Population	**Sample**
$P = 0.75$	$p = 50/60 = 0.83$
$Q = 1 - \text{P}$	$n = 60$
$=1 - 0.75 = 0.25$	

Step 2: Framing the hypothesis

$$H_0 : P \le 0.75$$

$$H_1 : P > 0.75$$

Step 3: Defining the test statistic

The test statistic to be evaluated is Z_c, and it can be defined as

$$Z_c = \frac{p-P}{SE(p)}, \text{ where } SE(p) = \sqrt{\frac{PQ}{n}}.$$

Step 4: Defining the significance level

Because the level of significance is not given, let us assume that $\alpha = 0.05$.

Since $H_1 : p > 0.75$, it implies that the test is a one-tailed (right-tailed) one (refer to Figure 13.2). According to the table, the value of $Z_t(\alpha) = Z_t(0.05) = 1.645$.

We see that, if Z_c of a sample statistic less than equal to 1.645, then we are 95% confident that H_0 is true ($P(Z \le 1.645) = 0.95$).

Step 5: Evaluate the value of $SE(P)$ and Z_c

$$SE(p) = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.75 \times 0.25}{60}} = \sqrt{0.003125} = 0.0559$$

$$Z_c = \frac{p - P}{SE(p)} = \frac{0.83 - 0.75}{0.0559} = 1.4311$$

Step 6: Statistical decisions

Since $Z_c = 1.4311$ lies in the acceptance region, ($Z_c < 1.645$) according to the decision rule, we accept H_0.

Step 7: Conclusion

At the 5% level of significance, we conclude that the population proportion, P be ≤ 0.75. The claim of the account is not correct.

13.11 Test for Equality of Two Proportions: Large Samples

Refer to Flowchart 13.5.

Example:

A survey of television audience in a big city revealed that 50 out of 200 males and 80 out of 250 females liked a particular nightly program. Test the hypothesis at the 5% level of significance whether there is a real difference of opinion about the program between male and female audiences.

Step 1: Given the values:

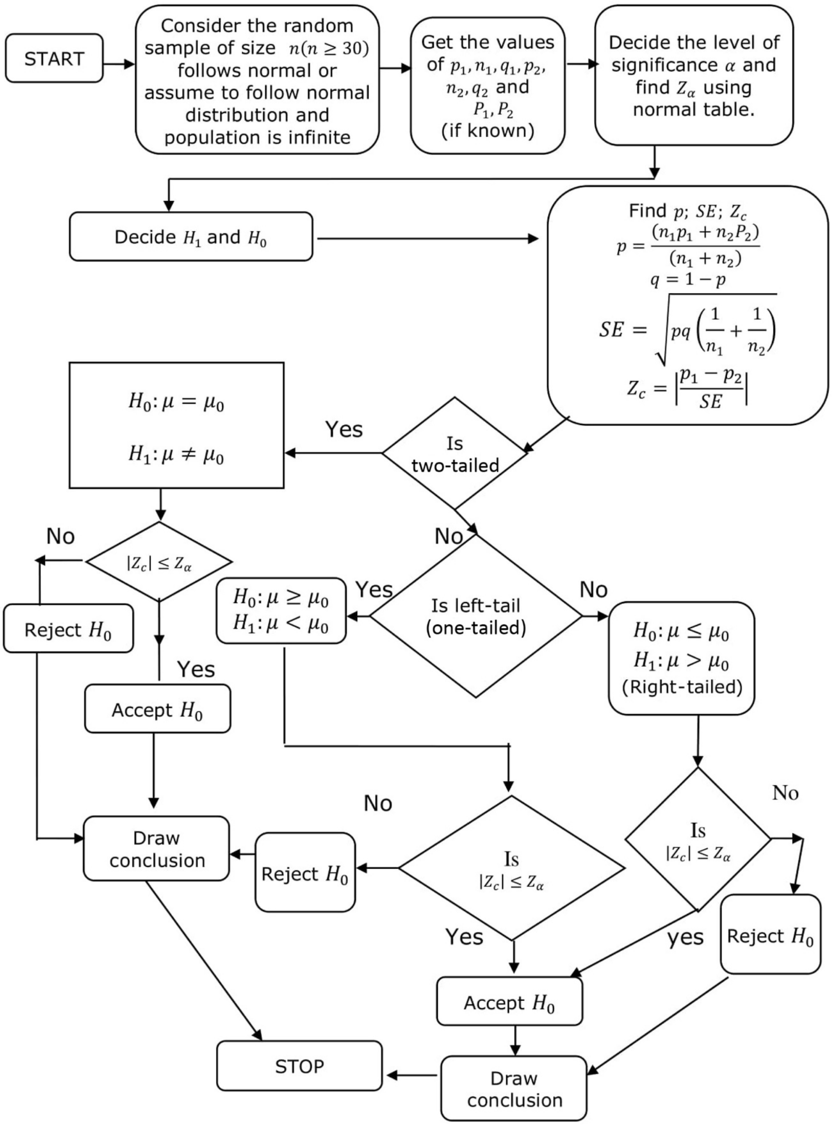
Sample 1 [Male]	**Sample 2 [Female]**
$n_1 = 200$	$n_2 = 250$
$p_1 = \frac{50}{200} = 0.25$	$p_2 = \frac{80}{250} = 0.32$
$q_1 = 1 - p_1 = 0.75$	$q_2 = 1 - p_2 = 0.68$

Let P_1 and P_2 be the two population proportions of the male and female, respectively.

Step 2: Framing the hypothesis

$$H_0 : P_1 = P_2$$

$$H_1 : P_1 \ne P_2$$


START
Consider the random sample of size $n (n \geq 30)$ follows normal or assume to follow normal distribution and population is infinite
Get the values of $p_1, n_1, q_1, p_2, n_2, q_2$ and P_1, P_2 (if known)
Decide the level of significance α and find Z_α using normal table.
Decide H_1 and H_0
Find p; SE; Z_c
$p = \frac{(n_1 p_1 + n_2 P_2)}{(n_1 + n_2)}$
$q = 1 - p$
$SE = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$
$Z_c = \left|\frac{p_1 - p_2}{SE}\right|$
Is two-tailed
Yes
$H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$
No
$|Z_c| \leq Z_\alpha$
No
Reject H_0
Yes
Accept H_0
Draw conclusion
Is left-tail (one-tailed)
Yes
$H_0: \mu \geq \mu_0$
$H_1: \mu < \mu_0$
No
$H_0: \mu \leq \mu_0$
$H_1: \mu > \mu_0$
(Right-tailed)
No
Is $|Z_c| \leq Z_\alpha$
Reject H_0
Yes
Is $|Z_c| \leq Z_\alpha$
No
yes
Reject H_0
Accept H_0
Draw conclusion
STOP

FLOWCHART 13.5
Test for equality of two population proportions (large sample).

Step 3: Defining the test statistic

The test statistic to be evaluated is Z_c, and it is defined as

$$Z_c = \frac{p_1 - p_2}{SE(p_1 - p_2)};\ \text{where } SE(p_1 - p_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\text{and } p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2};\ q = 1 - p.$$

Step 4: Defining the significance level

Here, decision sets the level of significance at $\alpha = 0.05$.

Since $H_1 : p_1 \neq p_2$, it implies that the test is a two-tailed one (refer to Figure 13.1). According to the table, the value of $Z_t(\alpha) = Z_t(0.05) = 1.96$.

We see that if Z_c of a sample statistic lies between –1.96 and +1.96, then we are 95% confidence that H_0 is true ($P(-1.96 \leq Z \leq 1.96) = 0.95$).

Step 5: Evaluate the value of $SE(p_1 - p_2)$ and Z_c

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{200 \times 0.25 + 250 \times 0.32}{200 + 250}$$

$$p = 0.2889; q = 1 - p = 0.7111$$

$$SE(p_1 – p_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{[0.2889 \times 0.7111] \times \left[\frac{1}{200} + \frac{1}{250}\right]}$$

$$SE(p_1 – p_2) = \sqrt{0.0019} = 0.0436$$

$$Z_c = \frac{p_1 - p_2}{SE(p_1 - p_2)} = \frac{0.25 - 0.32}{0.0436}; Z_c = -1.6055; |z_c| = |1.6055| = 1.6055$$

NOTE: Whenever $H_0 : P_1 - P_2 \neq 0$, then Z_c should be evaluated using a modified formula:

$$Z_c = \frac{[p_1 - p_2] - [P_1 - P_2]}{SE(p_1 - p_2)}$$

Step 6: Statistical decisions

Since $|z_c| = 1.6055$ lies in the acceptance region $|z| > 1.96$, according to the decision rule, we accept H_0.

Step 7: Conclusion

At the 5% level of significance, we conclude that there is no significant difference between the two population proportions.

Example:

A company is considering two different TV advertisements (ads) for the promotion of a new product. The CEO believes that ad A is more effective than ad B, so two test markets with virtually identical consumers are selected. Ad A is used in one area and ad B is used in another area. In a random sample of 60 customers who saw the ad A, 18 tried the product. In a random sample of 100 customers who saw ad B, 22 tried the product. Does this mean that ad A is more effective than ad B, if a 5% level of significance is used?

Step 1: Given the data:

Sample 1 [Ad A]	Sample 2 [Ad B]
$p_1 = 18/60 = 0.3$	$p_2 = 22/100 = 0.22$
$q_1 = 1 - p_1 = 0.7$	$q_2 = 1 - p_2 = 0.78$
$n_1 = 60$	$n_2 = 100$

Let P_1 and P_2 be the 2 proportions of the Population 1 and Population 2, respectively.

Step 2: Framing the hypothesis

$$H_0: P_1 \leq P_2$$

$$H_1: P_1 > P_2$$

Step 3: Defining the test statistic

The test statistic to be evaluated is Z_c, and it is defined as

$$Z_c = \frac{p_1 - p_2}{SE(p_1 - p_2)}; \text{ where } SE(p_1 - p_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\text{and } p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}; q = 1 - p.$$

Step 4: Defining the significance level

Hence, the decision has set the level of significance at $\alpha = 0.05$.

Since $H_{1:}$ $P_1 > P_2$ implies that the test is a one-tailed (right-tailed) one (refer to Figure 13.2), according to the table, the value of $Z_t(\alpha) = Z_t(0.05) = 1.645$.

We see that if Z_c of a sample statistic satisfies the condition $Z_c \leq 1.645$, we are 95% confident that H_0 is true ($P(Z_c \leq 1.645) = 0.95$).

Step 5: Evaluate the value of SE($p_1 - p_2$) and Z_c.

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{0.3 \times 60 + 0.22 \times 100}{60 + 100} = 0.25; q = 1 - 0.25 = 0.75$$

$$\text{SE}(p_1 - p_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{[.25 \times .75]\left(\frac{1}{60} + \frac{1}{100}\right)} = \sqrt{0.005} = 0.0707$$

$$\text{SE}(p_1 - p_2) = 0.0707$$

$$Z_c = \frac{p_1 – p_2}{SE(p_1 - p_2)} = \frac{0.3 - 0.22}{0.0707} = 1.1315; Z_c = 1.1315$$

Step 6: Statistical decisions

Since $Z_c = 1.1315$ lies in the acceptance region $Z \leq 1.645$, according to the decision rule, we accept H_0.

Step 7: Conclusion

At 5% level of significance, we conclude that the Ad A is not more effective than Ad B.

Example:

In a certain city, 125 men in a sample of 500 were found to be smokers. In another city, the numbers of smokers were 375 in a random sample of 1000. Does it indicate that there is a greater population of smokers in the second city than in the first?

Step 1: Given the data:

Sample 1	**Sample 2**
$p_1 = 125/500 = 0.25$	$p_2 = 375/1000 = 0.375$
$q_1 = 1 - p_1 = 0.75$	$q_2 = 1 - p_2 = 0.625$
$n_1 = 500$	$n_2 = 1000$

Let P_1 and P_2 be the two proportions of the population: Population 1 and Population 2, respectively.

Step 2: Framing the hypothesis

$$\mathrm{H}_0\text{: } P_1 \geq P_2$$

$$\mathrm{H}_1\text{: } P_1 < P_2$$

Step 3: Defining the test statistic

The test statistic to be evaluated is Z_c, and it is defined as

$$Z_c = \frac{p_1 - p_2}{SE(p_1 - p_2)}\text{; where } SE(p_1 - p_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\text{and } p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}\text{; } q = 1 - p.$$

Step 4: Defining the significance level

Because the level of significance at α is not given, let us assume that $\alpha = 0.05$.

Since H$_1$: $P_1 < P_2$ implies that the test is a one-tailed (left-tailed) one (refer to Figure 13.3), according to the table, the value of $Z_t(\alpha) = Z_t(0.05) = -1.645$. We see that if Z_c of a sample statistic satisfies the condition $Z_c >= -1.645$, we are 95% confident that H$_0$ is true ($P(Z_c >= 1.645) = 0.95$).

Step 5: Evaluate the value of $SE(p_1 – p_2)$ and $Z_{c.}$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{0.25 \times 500 + 0.375 \times 1000}{500 + 1000} = 0.333;\ q = 1 – .333 = 0.667$$

$$SE(p_1 – p_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{[.333 \times .667]\left(\frac{1}{500} + \frac{1}{1000}\right)} = \sqrt{0.000666} = 0.0258$$

$$SE(p_1 – p_2) = 0.0258$$

$$Z_c = \frac{p_1 - p_2}{SE(p_1 - p_2)} = \frac{0.25 - 0.375}{0.0258} = -4.845;\ Z_c = -4.845$$

Step 6: Statistical decisions

Since $Z_c = -4.845$ lies in the critical region $Z \leq -1.645$, according to the decision rule, we reject H_0.

Step 7: Conclusion

At 5% level of significance, we conclude that the proportion of smokers is more in the second city than in the first city.

13.12 Test for Equality of Two Standard Deviations: Large Samples

Example:

Random samples drawn from two different populations gave the following data relating to the heights of adult males:

	Sample 1	Sample 2
Average Height	67.42 inches	67.25 inches
SD	2.58 inches	2.50 inches
Size	1000	1200

Is the difference between the SDs significant?

Step 1: Given the data:

Sample 1	Sample 2
$X_1 = 67.42$	$X_2 = 67.25$
$s_1 = 2.58$	$s_2 = 2.50$
$n_1 = 1000$	$n_2 = 1200$

Let σ_1 and σ_2 are the two SDs of the population 1 and population 2, respectively.

Step 2: Framing the hypothesis

$$H_0: \sigma_1 = \sigma_2;\ H_1: \sigma_1 \neq \sigma_2$$

Step 3: Defining the test statistic

Because the samples are large, we make use of the statistics Z.

The test statistic Z_c can be computed using the relation

$$Z_c = \frac{s_1 - s_2}{SE[s_1 - s_2]}$$

$$\text{where } SE[s_1 - s_2] = \sqrt{\frac{s1^2}{2n_1} + \frac{s2^2}{2n_2}}$$

Step 4: Defining the significance level

Because the level of significance at α is not given, let us assume that $\alpha = 0.05$.

Since H$_1$: $\sigma_1 \neq \sigma_2$ implies that the test is a two-tailed one, according to the table, the critical value of $Z_t(\alpha) = Z_t(0.05) = 1.96$.

We see that if Z_c of the sample statistic lies between –1.96 and +1.96, so we are 95% confident that H$_0$ is true ($P(-1.96 \leq Z \leq 1.96) = 0.95$).

Step 5: Evaluate $SE(s_1–s_2)$ and Z_c.

$$SE[s_1 - s_2] = \sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}} = \sqrt{\frac{2.58^2}{2(1000)} + \frac{2.50^2}{2(1200)}} = \sqrt{0.0059} = 0.0768$$

$$SE[s_1 - s_2] = 0.0768$$

$$Z_c = \frac{s_1 - s_2}{SE[s_1 - s_2]}$$

$$Z_c = \frac{2.58 - 2.50}{0.0768} = 1.0417$$

$$Z_c = 1.0417$$

Step 6: Statistical Decisions

Since $Z_c = 1.0417$, which lies in the acceptance region ($1.0147 \leq 1.967$), according to the decision rule, we accept H$_0$.

Step 7: Conclusion

We conclude that at the 5% level of significance, there is no significant difference between the SDs of the two populations.

13.13 Student's t-Distribution

Consider a population with mean, μ, and the variance, $\sigma^2_,$ follows normal distribution. Select m number of small samples of size, n. Let it be $(S_1, n), (S_2, n), \ldots, (S_m, n)$. Then find the means of each sample. Let it be $x_1, x_2, \ldots x_n$. By considering all these m values, construct the discrete distribution with frequency. The resulting distribution is known as a student's t-distribution.

Then the student's t-statistic can be defined as, $t = \dfrac{\bar{x}_b - \mu}{\left[{}^{S}\!/\!\sqrt{n}\right]}$

Where $\bar{x}$ = sample mean $= \left(\frac{1}{n}\right)\sum_{i=1}^{n} \Sigma x_i$

s^2 = sample SD $= \left(\frac{1}{n}\right)\sum_{i=1}^{n} \Sigma(x_i - \bar{x})^2$

s^2 is an unbiased estimate of the population variance σ^2.

Then the t-distribution with $(n–1)$ degrees of freedom can be given by,

$$f(t) = c\left(1 + \frac{t^2}{v}\right)^{-\left(\frac{v+1}{2}\right)}$$

Where $v = n - 1$, degrees of freedom
c = is a constant.

The value of c can be evaluated using the definite integral $\int_{-\infty}^{+\infty} f(t)dt = 1$.

NOTE 1:

$$s^2 = \left(\frac{1}{n}\right)\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{13.1}$$

$$S^2 = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{13.2}$$

Equations (13.1) and (13.2) imply that, $(n - 1)S^2 = (n)s^2$, then

$$\frac{S^2}{n} = \frac{s^2}{n-1}; \frac{S}{\sqrt{n}} = \frac{s}{\sqrt{n-1}}$$

NOTE 2: When $n \to \infty$, the small sample becomes the large sample. Then $S = s$.

Therefore, $t = \dfrac{\overline{x} - \mu}{\left(S/\sqrt{n}\right)} = \dfrac{\overline{x} - \mu}{\left(s/\sqrt{n-1}\right)}$ is the test statistic.

The nature of the distribution of 't' was first introduced and discussed by William Sealy Gosset. Gosset published the research work using the pseudonym 'Student'. Hence this t-distribution is usually referred to as 'student's distribution'.

13.14 Properties of t-Distribution

1. It is defined as,

   $$f(t) = c\left(1 + \frac{t^2}{v}\right)^{-\left(\frac{v+1}{2}\right)}$$

   Where t = $t = \dfrac{\overline{x} - \mu}{\left(S/\sqrt{n}\right)} = \dfrac{\overline{x} - \mu}{\left(s/\sqrt{n-1}\right)}$

   $-\infty < t < +\infty$; c is a constant; and $v = n–1$ (degrees of freedom)

2. $\int_{-\infty}^{\infty} f(t)dt = 1$

3. The mean value is 0.

4. In general, if the value of variance is more than 1, it approaches the value 1 as $n \to \infty$.

5. Variance $= \dfrac{v}{v-2}; v > 2.$
6. In general, the t-distribution is less peaked at the centre and higher in the tails than the normal distribution.
7. The t-distribution approaches the normal distribution as $n \to \infty$. Assumptions related to t-distribution:
    - The sample should be a small sample ($n < 30$).
    - The sample is selected randomly.
    - The population is normal.
    - The SD of the population is not known.

13.15 Test for the Specified Mean: Small Sample

Refer to Flowchart 13.6.

Example:

Given a sample mean of 83, a sample SD of 12.5, and a sample size of 22, test the hypothesis that the value of the population mean is 70 against the alternative that it is more than 70. Use the 0.05 significance level.

Step 1: Given the data:

Sample	**Population**
$\bar{x} = 83$	$\mu = 70$
$s = 12.5$	
$n = 22$	
$v = 22 - 1 = 21$	

Step 2: Framing the hypothesis

$$H_0 : \mu = 70; H_1 : \mu > 70$$

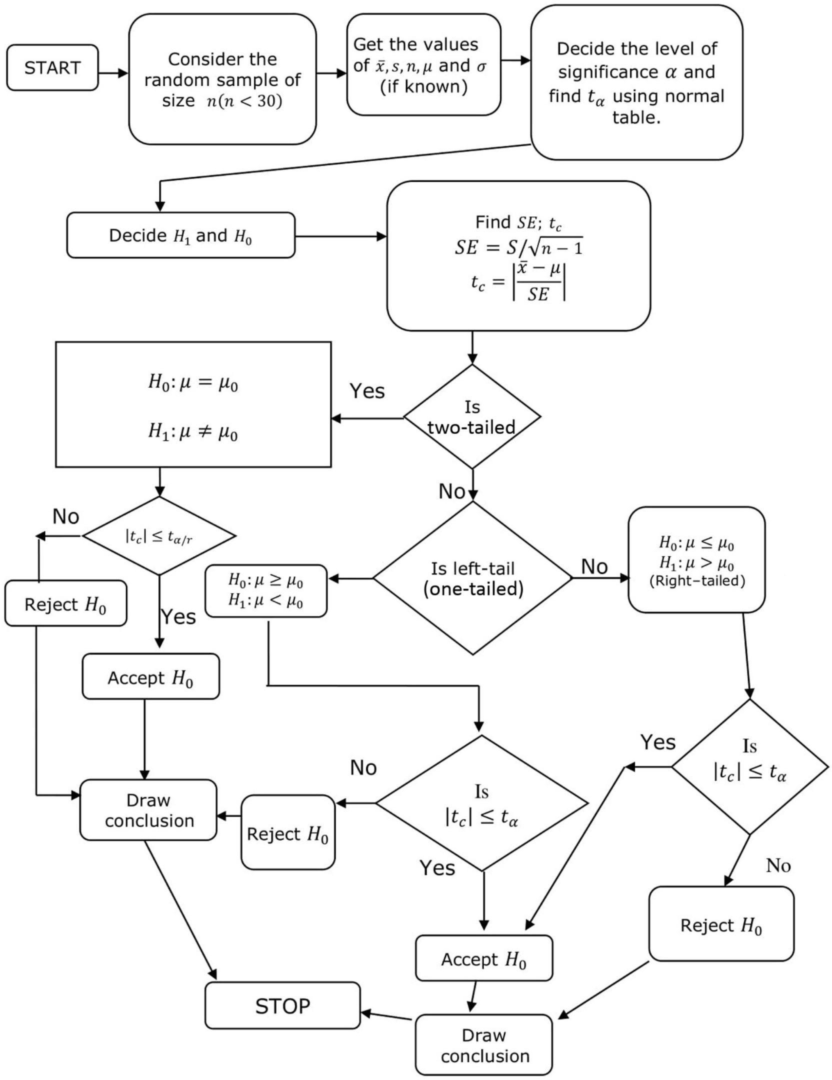
Step 3: Defining the test statistic

Since σ is unknown and n = 22 < 30, it implies that the given sample is small. The test statistic to be evaluated is t_c. (t-calculated value), and it is defined as,

$$t_c = \frac{\bar{x} - \mu}{SE(\bar{x})}, \text{ where } SE(\bar{x}) \frac{s}{\sqrt{n-1}}$$

Step 4: Defining the significance level

Here, decision sets the level of significance at $\alpha = 0.05$. $\because H_1 : \mu > 70$, implies that the test is a one-side (right-tailed) test. According to the table, the value of $t_t(\alpha, v) = t_t(0.05, 21) = 1.721$.



FLOWCHART 13.6
Test for a specifies mean (small sample).

We see that if t_c of the sample statistic satisfied the condition $|t_c| \leq t_t$, then we are 95% confident that H_0 is true $[P(|t_c| \leq 1.721) = 0.95]$.

Step 5: Evaluate $SE(\bar{x})$and t_c.

$$SE(\bar{x}) = \frac{s}{\sqrt{n-1}} = \frac{12.5}{\sqrt{22-1}} = \frac{12.5}{\sqrt{21}} = 2.7277.\ t_c = \frac{\bar{x}-\mu}{SE(\bar{x})} = \frac{83-70}{2.7277} = \frac{13}{2.7277}$$

$$t_c = 4.7659$$

Step 6: Statistical decisions

Since $t_c = 4.7652$ lies in the critical region, $t_t > 1.721$, according to the decision rule, we reject H_0.

Step 7: Conclusion

We conclude that at the 5% level of significance with 21 *df*, the population mean is more than 70.

Example:

A certain medicine administered to each of 10 patients resulted in the following increases in the blood pressure (BP)" 8, 8, 7, 5, 4, 1, 0, 0, –1, –1. Can it be concluded that the medicine was responsible for the increase in BP?

Step 1: Based on the given the data, construct the frequency distribution table.

Increase in BP, X	Frequency, f
–1	2
0	2
1	1
4	1
5	1
7	1
8	2
Total	10

Hence, evaluate the value of mean and SD

X	f	fx	x²	fx²
–1	2	–2	1	2
0	2	0	0	0
1	1	1	1	1
4	1	4	16	16
5	1	5	25	25
7	1	7	49	49
8	2	16	64	128
Total	10	31		221

$$\text{Here, } \sum_{i=1}^{7} f_i = 10$$

$$\bar{x} = \frac{\sum_{i=1}^{7} f_i x_i}{\sum_{i=1}^{7} f_i} = \frac{31}{10} = 3.1$$

$$s^2 = \frac{\sum_{i=1}^{7} f_i {x_i}^2}{\sum_{i=1}^{7} f_i} - \left(\frac{\sum_{i=1}^{7} f_i x_i}{\sum_{i=1}^{7} f_i} \right)^2$$

$$s^2 = \frac{221}{10} - \left(\frac{31}{10} \right)^2 = 12.49$$

$$s = \sqrt{12.49} = 3.5341$$

$$\bar{X} = 3.1$$

$$s = 3.5341$$

$$n = 10$$

$$df = n - 1 = 10 - 1 = 9$$

Step 2: Framing the hypothesis

H$_0$: Medicine is not responsible for the increase in BP.

H$_1$: Medicine is responsible for the increase in BP.

Step 3: Defining the test statistic

Since σ is unknown and $n = 10 < 30$, the sample is a small sample. The statistic to be evaluated is t_c. It is defined as,

$$t_c = \frac{\bar{x} - \mu}{SE(\bar{x})};\text{ where } SE(\bar{x}) = \frac{s}{\sqrt{n-1}}.$$

Step 4: Defining the significance level

Here, decision sets the level of significance at $\alpha = 0.05$. $\because H_1$, implies that the test is a two-side (two-tailed) test. According to the table, the value of $t_t(\alpha, \nu) = t_t(0.05, 9) = 2.262$.

We see that if t_c of a sample statistic satisfies the condition $|t_c| \leq 2.262$, then we are 95% confident that H$_0$ is true ($P(|t_c| \leq 2.262) = 0.95$).

Step 5: Evaluate $SE(\bar{x})$ and t_c

$$SE(\bar{x}) = \frac{s}{\sqrt{n-1}} = \frac{3.5341}{\sqrt{10-1}} = \frac{3.5341}{3}$$

$$SE(\bar{x}) = 1.178$$

$$t_c = \frac{\bar{x} - \mu}{SE(\bar{x})} = \frac{3.1 - 0}{1.178}$$

$$t_c = 2.6316$$

Step 6: Statistical decisions

Since $t_c = 2.6316$ lies in the critical region $|t_c| > 2.262$, according to the decision rule, we reject H$_0$.

Step 7: Conclusion

We conclude that at 5% level of significance with 9 *df*, the medicine was responsible for the increase in BP.

Example:

A salesman is expected to effect mean sales of \$3500. A sample test revealed that a particular salesman had made the following sales: \$2000, \$3000, \$5200, \$3400, \$2500, and \$3700. Using 1% level of significance, can we conclude whether this work is below standard?

Step 1: Based on the given data, evaluate the mean and SD.

Here, $n = 6$

$$\text{Mean} = \frac{\sum_{i=1}^{6} x_i}{n} = \frac{19800}{6} = 3300$$

Sales (X)	$\bar{x} = 3300$ $x - \bar{x}$	$(x - \bar{x})^2$
\$2000	–1300	169×10^4
\$3000	–300	9×10^4
\$5200	1900	361×10^4
\$3400	100	1×10^4
\$2500	–800	64×10^4
\$3700	400	16×10^4
$\Sigma(x - \bar{x})^2 = 620 \times 10^4$		

$$\bar{x} = \$3300$$

$$s^2 = \sum_{i=1}^{6} \frac{(x_i - \bar{x})^2}{n} = \frac{620 * 10^4}{6}$$

$$s^2 = 1033333.333;$$

$$s = \sqrt{1033333.333};\ s = 1016.5301$$

Hence,

$$\bar{x} = \$3300$$

$$s = \$1016.5301$$

$$n = 6$$

$$df = n–1 = 6 - 1 = 5$$

Step 2: Framing the hypothesis

$$\text{H}_0 : \mu \geq \text{Rs. } 3500$$

$$\text{H}_1 : \mu < \text{Rs. } 3500$$

Step 3: Testing the test statistic

Since σ is unknown and $n = 6 < 30$, the given sample is said to be small sample. The statistic to be evaluated is $t_{c.}$

$$t_c = \frac{\bar{x} - \mu}{SE(\bar{x})}; \text{ where } SE(\bar{x}) = \frac{s}{\sqrt{n-1}}.$$

Step 4: Defining the significance level

The level of significance α is given as 1%. The alternative hypothesis H_1: $\mu < 3500$ implies that it is a one-tailed (left-tailed) test.

According to the table, the critical value of $t_t(\alpha,\nu) = t_t(0.01,5) = 3.365$. We see that if t_c of a sample statistic satisfies the condition $t_c \geq -3.365$, then we are 99% confident that H_0 is true ($P(t_c \geq -3.365) = 0.99$).

Step 5: Evaluate $SE(\bar{x})$ and t_c.

$$SE(\bar{x}) = \frac{s}{\sqrt{n-1}} = \frac{1016.5301}{\sqrt{6-1}} = 454.6061$$

$$t_c = \frac{\bar{x} - \mu}{SE(\bar{x})} = \frac{3300 - 3500}{454.6061}$$

$$t_c = -0.4399$$

Step 6: Statistical decisions

Since $t_c = -0.4399$ lies in the acceptance region, according to the decision rule, we accept H_0.

Step 7: Conclusion

We conclude that at 1% level of significance with 5 *df*, the salesman is up to the standards.

13.16 Test for Equality of Two Population Means: Small Samples

(σ_1 and σ_2 are unknown.)

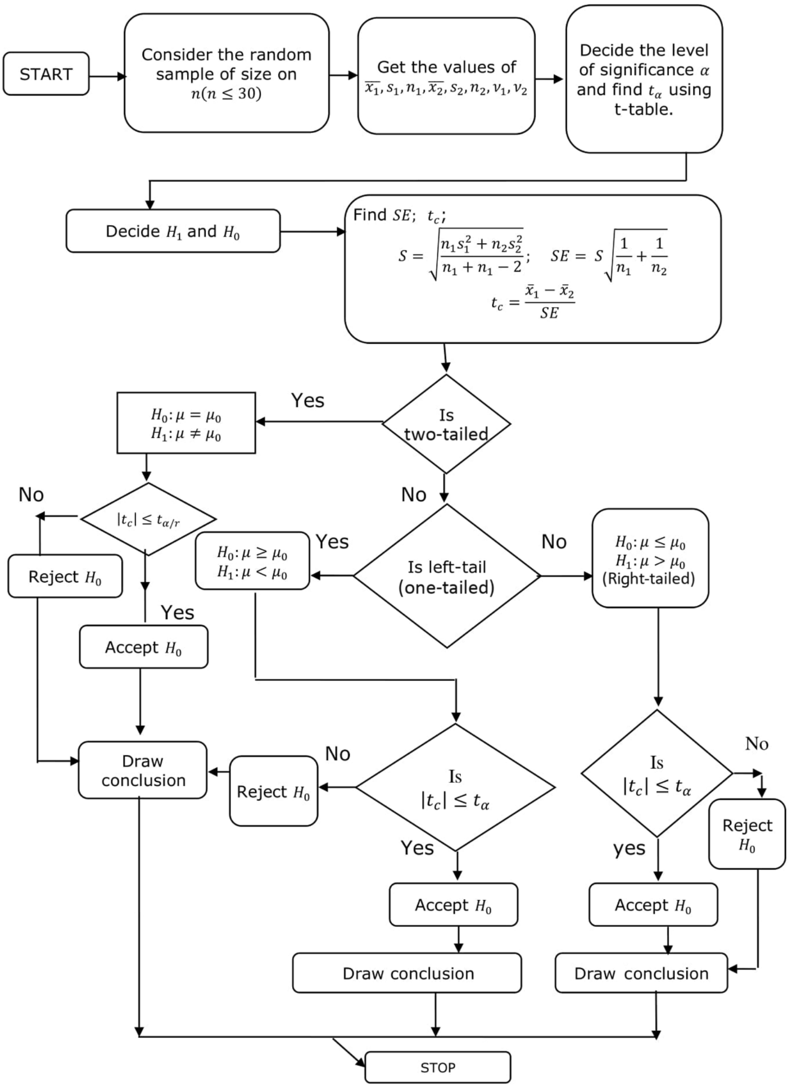
Refer to Flowchart 13.7.

Example:

Two random samples gave the following results:

Sample	Size	Sample Mean	Sum of Squares of Deviations from Mean
1	10	15	90
2	12	14	108

Assuming normal population, test for the equality of population variances at 5% level of significance.


START
Consider the random sample of size on $n(n \leq 30)$
Get the values of $\overline{x_1}, s_1, n_1, \overline{x_2}, s_2, n_2, v_1, v_2$
Decide the level of significance α and find t_α using t-table.
Decide H_1 and H_0
Find SE; t_c;
$S = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_1 - 2}}$; $SE = S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
$t_c = \frac{\bar{x}_1 - \bar{x}_2}{SE}$
Is two-tailed
Yes
$H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$
No
$|t_c| \leq t_{\alpha/r}$
No
Reject H_0
Yes
Accept H_0
Draw conclusion
Is left-tail (one-tailed)
Yes
$H_0: \mu \geq \mu_0$
$H_1: \mu < \mu_0$
No
$H_0: \mu \leq \mu_0$
$H_1: \mu > \mu_0$
(Right-tailed)
Is $|t_c| \leq t_\alpha$
No
Reject H_0
Yes
Accept H_0
Draw conclusion
Is $|t_c| \leq t_\alpha$
No
Reject H_0
yes
Accept H_0
Draw conclusion
STOP

FLOWCHART 13.7
Test for equality of two population means (small sample).

Step 1: Based on the given data, find the SDs of the two samples.

Sample 1	Sample 2
$\overline{x}_1 = 15$	$\overline{x}_2 = 14$
$n_1 = 10$	$n_2 = 12$
$s_1 = \sqrt{\frac{90}{10}} = 3$	$s_2 = \sqrt{\frac{108}{12}} = 3$

Let μ_1 and μ_2 be the means of two populations.
$v = df = [10 - 1] + [12 - 1] = 20$

Step 2: Framing the hypothesis

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Test for difference of two population means.

Step 3: Defining the test statistic

Since σ_1 and σ_2 are unknown and the two samples are small (n_1, $n_2 < 30$), then test statistic to be evaluated is t_c.

$$t_c = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)}$$

$$\text{Where } SE(\bar{x}_1 - \bar{x}_2) = s_c\sqrt{\tfrac{1}{n_1} + \tfrac{1}{n_2}};\ s_c = \sqrt{\tfrac{n_1 s_1{}^2 + n_2 s_2{}^2}{n_1 + n_2 - 2}}$$

Step 4: Defining the significance level

Because the level of significance α is not given, let us assume that $\alpha = 0.05$. The alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ implies that it is a two-tailed test.

According to the table, the critical value of $t(\alpha,\nu) = t_t(0.05,20) = 2.086$. We see that if t_c of the sample statistic satisfies the condition $|t_t| \leq 2.086$, then we are 95% confident that H_0 is true ($P(|t_t| \leq 2.086) = 0.95$).

Step 5: Evaluate $SE(\bar{x}_1 - \bar{x}_2)$ and t_c

$$SE(\bar{x}_1 - \bar{x}_2) = s_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s_c = \sqrt{\frac{n_1 s_1{}^2 + n_2 s_2{}^2}{n_1 + n_2 - 2}} = \sqrt{\frac{10(9) + 12(9)}{10 + 12 - 2}} = \sqrt{\frac{198}{20}} = 3.1464.$$

$$s_c = 3.1464$$

$$SE(\bar{x}_1 - \bar{x}_2) = 3.1464\sqrt{\frac{1}{10} + \frac{1}{12}} = 3.1464 \times 0.4282$$

$$SE(\bar{x}_1 - \bar{x}_2) = 1.3473;\ t_c = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)} = \frac{15 - 14}{1.3473} = \frac{1}{1.3473};\ t_c = 0.7422$$

Step 6: Statistical decisions

Since $t_c = 0.7422$; lies in the acceptance region, ($t_c \leq t_t$; $0.7422 < 2.086$), according to the decision rule, we accept H_0.

Step 7: Conclusion

We conclude that at 5% level of significance with 20 *df*, there is no significant difference between the two population means.

Example:

A group of 5 patients treated with medicine A weigh 42, 39, 48 60, and 41 kg, respectively. A second group of 7 patients from the same hospital treated with medicine B weigh 38, 42, 56, 64, 68, 69, and 62 kg respectively. Do you agree with the claim that medicine B increases the weight significantly? Test with 5% level of significance.

Step 1: Based on the given data, find the means and SDs of the two samples.

Sample 1: Medicine A

$n = 5$

$$\bar{x}_1 = a + \frac{\sum_{i=1}^{5} Y_i}{n}$$

	$a = 48$	
X (Weight [kg])	**$y = x - a$**	**Y^2**
42	−6	36
39	−9	81
48	0	0
60	12	144
41	−7	49
Total	−10	310

$$\bar{x}_1 = 48 + \left(\frac{-10}{5}\right) = 46.$$

$$s_1^2 = \frac{\sum_{i=1}^{5} Y_i^2}{n} - \left(\frac{\sum_{i=1}^{5} Y_i}{n}\right)^2$$

$$= \frac{310}{5} - \left(\frac{-10}{5}\right)^2$$

$$= \frac{310}{5} - 4 = \frac{290}{5}$$

$$s_1 = \sqrt{\frac{290}{5}} = \sqrt{58} = 7.6158$$

NOTE: Use of assumed mean method.

Sample 2: Medicine B

X (Weight [kg])	$a = 50$ $y = x - a$	Y^2
38	−12	144
42	−8	64
56	6	36
64	14	196
68	18	324
69	19	361
62	12	144
Total	49	1269

$$n = 7$$

$$\bar{x}_2 = a + \frac{\sum_{i=1}^{7} Y_i}{n}$$

$$\bar{x}_2 = 50 + \left(\frac{49}{7}\right) = 57$$

$$s_2^2 = \frac{\sum_{i=1}^{7} Y_i^2}{n} - \left(\frac{\sum_{i=1}^{7} Y_i}{n}\right)^2$$

$$= \frac{1269}{7} - \left(\frac{49}{7}\right)^2 = 132.2857$$

$$s_2 = \sqrt{132.2857} = 11.5016$$

$$v = df = [5 - 1] + [7 - 1] = 10$$

Step 2: Framing the hypothesis

$$\begin{matrix} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{matrix} \quad [\text{Left-tailed}]$$

Step 3: Defining the test statistic

Since σ_1 and σ_2 are unknown, and the samples are small, the test statistic to be evaluated is t_c.

It is defined as $t_c = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)}$

where, $SE(\bar{x}_1 - \bar{x}_2) = s_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$; $s_c = \sqrt{\frac{n_1 s_1{}^2 + n_2 s_2{}^2}{n_1 + n_2 - 2}}$

Step 4: Defining the significance level

Here, the level of significance α is given as α = 0.05. The alternative hypothesis H$_1$: $\mu_1 < \mu_2$, implies that it is a one-tailed (left-tailed) test. According to the table, the critical value of $t_t(\alpha,\nu) = t_t(0.05,10) = 1.812$. We see that if t_c of the test statistic satisfied the condition $t_c \leq -1.812$, then we are 95% confident that H$_0$ is true ($P(t_c \geq -1.812) = 0.95$).

Step 5: Evaluate $SE(\overline{x}_1 - \overline{x}_2)$ and t_c.

$$s_c = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{5(58) + 7(132.2857)}{5 + 7 - 2}}$$

$$S_c = \sqrt{121.60} = 11.0272$$

$$SE(\overline{x}_1 - \overline{x}_2) = S_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 11.0272\sqrt{\frac{1}{5} + \frac{1}{7}}$$

$$= 11.0272(0.5855) = 6.4564$$

$$SE(\overline{x}_1 - \overline{x}_2) = 6.4564.$$

$$t_c = \frac{\overline{x}_1 - \overline{x}_2}{SE(\overline{x}_1 - \overline{x}_2)}$$

$$= \frac{46 - 57}{6.4564}$$

$$t_c = -1.7037$$

Step 6: Statistical decisions

Since $t_c = -1.7033$, which lies in the acceptance region ($-1.812 \leq -1.7037$), we accept H$_0$ according to the decision rule.

NOTE: $|t_c| = |-1.7037| = 1.7037$ and the $|t_t| = |-1.812| = 1.812$; testing condition $|t_c| \leq |t_t|$. Clearly $1.7033 \leq 1.812$; according the decision rule we accept H$_0$.

Step 7: Conclusion

We conclude that at 5% level of significance with 10 *df*, medicine B is not increasing weight significantly.

Example:

Two samples of sodium vapour bulbs were tested for length of life, and the results are as follows:

	Size of Sample	Sample Mean (Hours)	Sample SD (Hours)
Type I	8	1234	36
Type II	7	1036	40

Is the difference in the means sufficient to generalize that type I is superior to type II with regard to length of life?

Step 1: Given the values:

Sample I	Sample II
(Type I)	(Type II)
$\bar{x}_1 = 1234$ hrs.,	$\bar{x}_2 = 1036$ hrs.,
$s_1 = 36$	$s_2 = 40$
$n_1 = 8$	$n_2 = 7$

Let μ_1 and μ_2 be the two population means.

$$v = df = [8 - 1] + [7 - 1] = 13$$

Step 2: Framing the hypothesis

$$H_0 : \mu_1 \le \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

Step 3: Selecting the test statistic

Because σ_1 and σ_1 are unknown and the two samples are small, the test statistic to be evaluated is t_c.

It is defined as $t_c = \dfrac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)}$

Where $SE(\bar{x}_1 - \bar{x}_2) = s_c\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$; where $s_c = \sqrt{\dfrac{n_1 s_1{}^2 + n_2 s_2{}^2}{n_1 + n_2 - 2}}$

Step 4: Defining the significance level

Since α is not given, let us assume that $\alpha = 0.05$. The alternative hypothesis H_1: $\mu_1 > \mu_2$ implies that it is a one-tailed (right-tailed) test.

According to the table, the critical value of $t_t(\alpha,v) = t_t(0.05, 13) = 1.771$. We see that if t_c of the sample statistic satisfies the condition $t_c \le 1.771$, then we are 95% confident that H_0 is true ($P(t_c \le 1.771) = 0.95$).

Step 5: Evaluate $SE(\bar{x}_1 - \bar{x}_2)$ and t_c.

$$s_c = \sqrt{\frac{n_1 s_1{}^2 + n_2 s_2{}^2}{n_1 + n_2 - 2}} = \sqrt{\frac{8(36)^2 + 7(40)^2}{8 + 7 - 2}}$$

$$s_c = \sqrt{1659.0769} = 40.7318$$

$$SE(\bar{x}_1 - \bar{x}_2) = s_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (40.7318)\sqrt{\frac{1}{8} + \frac{1}{7}}$$

$$SE(\bar{x}_1 - \bar{x}_2) = 21.0807$$

$$t_c = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)} = \frac{1234 - 1036}{21.0807} = 9.3925$$

Step 6: Statistical decisions

Since $t_c = 9.3925$ lies in the critical region, that is $t_c = 9.3925 > 1.771$, according to the decision rule, we reject H_0.

Step 7: Conclusion

We conclude that at 5% level of significance with 13 *df*, type I bulb is superior to type II bulb regarding length of life.

13.17 Paired t-Test for Difference of Mean

Refer to Flowchart 13.8.

Example:

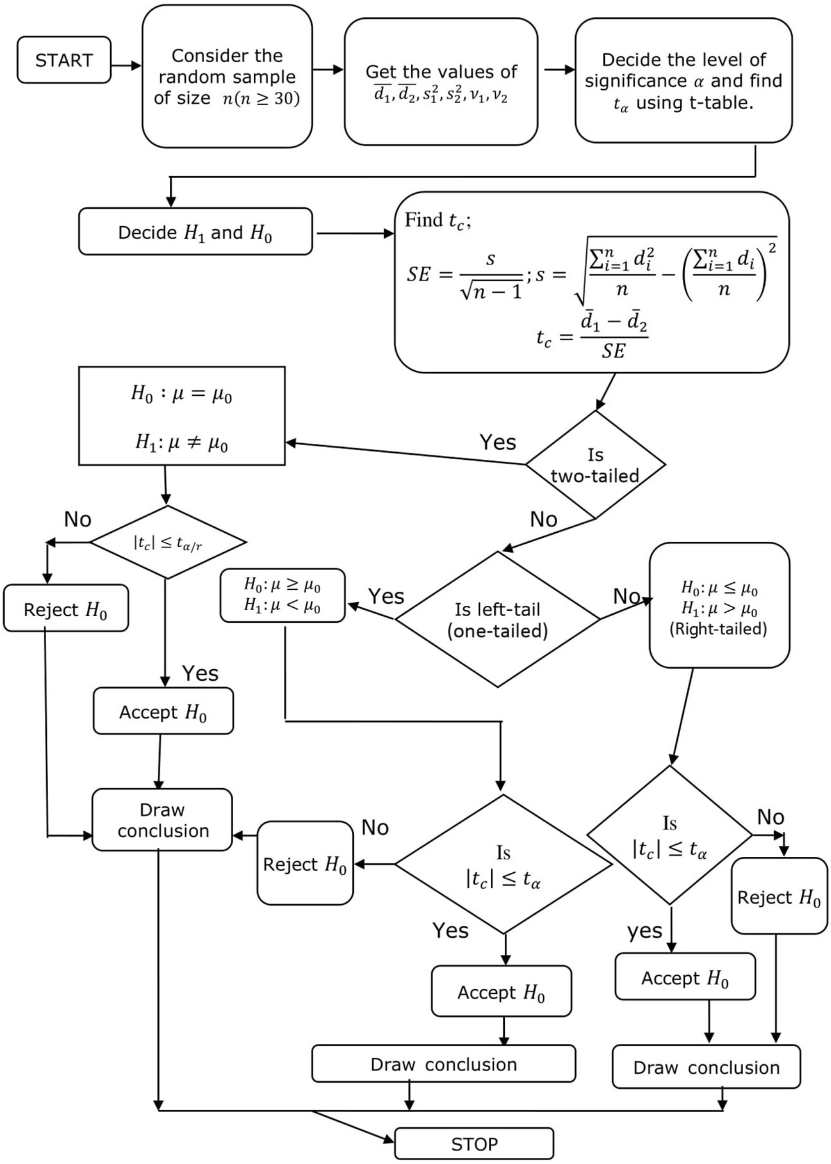
A manufacturer, who wishes to increase employee production, selects a department with 12 employees for an experiment. The manufacturer tries to improve the working conditions in this department through renovation and employee incentives. The following table shows the mean number of items produced per day by the employees 1 month before and 1 month after the changes are made. Verify whether there is any significant difference in the mean production before and after the changes made with a level of significance 5%.

Mean number of items produced per day

Employee	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Before	75	61	62	68	58	70	59	79	68	80	64	72
After	82	70	74	80	65	80	70	88	77	90	75	87

Step 1: Based on the given data, find the mean difference $\bar{d}$ and SDs.

Mean Production per Day Before Change (x)	**Mean Production per Day After Change (y)**	$d = y - x$	d^2
75	82	7	49
61	70	9	81
62	74	12	144
68	80	12	144
58	65	7	49
70	80	10	100
59	70	11	121
79	88	9	81
68	77	9	81
80	90	10	100
64	75	11	121
75	87	12	144
Total		119	1215


START
Consider the random sample of size $n(n \geq 30)$
Get the values of $\overline{d_1}, \overline{d_2}, s_1^2, s_2^2, \nu_1, \nu_2$
Decide the level of significance α and find t_α using t-table.
Decide H_1 and H_0
Find t_c;
$SE = \frac{s}{\sqrt{n-1}}; s = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n} - \left(\frac{\sum_{i=1}^{n} d_i}{n}\right)^2}$
$t_c = \frac{\bar{d}_1 - \bar{d}_2}{SE}$
$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$
Yes
Is two-tailed
No
No
$|t_c| \leq t_{\alpha/r}$
Reject H_0
Yes
Accept H_0
Draw conclusion
$H_0 : \mu \geq \mu_0$
$H_1 : \mu < \mu_0$
Yes
Is left-tail (one-tailed)
No
$H_0 : \mu \leq \mu_0$
$H_1 : \mu > \mu_0$
(Right-tailed)
No
Reject H_0
Is $|t_c| \leq t_\alpha$
Yes
Accept H_0
Draw conclusion
Is $|t_c| \leq t_\alpha$
No
Reject H_0
yes
Accept H_0
Draw conclusion
STOP

FLOWCHART 13.8
Paired *t*-test for the difference of means (small sample).

$$n = 12.$$

$$\text{Mean} = \bar{d} = \frac{\sum_{i=1}^{12} d_i}{n} = \frac{119}{12} = 9.9167$$

$$s=\sqrt{\frac{\sum_{i=1}^{12} d_i^2}{n}-\left(\frac{\sum_{i=1}^{12} d_i}{n}\right)^2}=\sqrt{\frac{1215}{12}-\left(\frac{119}{12}\right)^2}=\sqrt{2.9097}=1.7058$$

$$\bar{d}=9.9167;\ s=1.7058;\ n=12;\ df=12-1=11.$$

Let μ_1 and μ_2 stand for the mean production of the population before change and after change, respectively.

Step 2: Framing the hypothesis

$$H_0:\mu_1=\mu_2$$

$$H_1:\mu_1<\mu_2$$

Step 3: Defining the test statistic

Since σ is unknown and the sample size is 12 (12 < 30), which is a small sample, the statistic to be evaluated is t_c.

It is defined as

$$t_c=\frac{\bar{d}}{SE(\bar{d})};\ \text{where } SE(\bar{d})=\frac{s}{\sqrt{n-1}}$$

Step 4: Defining the significance level

The level of significance $\alpha = 0.05$. The alternative hypothesis H_1: $\mu_1 \neq \mu_2$ implies that it is a two-tailed test.

According to the table, the critical value of $t_t(\alpha,v) = t_t(0.05,11) = 2.201$. We see that if t_c of the sample statistic satisfies the condition $t_c \leq 2.201$, then we are 95% confident that H_0 is true ($P(t_c \leq 2.201) = 0.95$).

Step 5: Evaluate $SE(\bar{d})$ and t_c.

$$SE(\bar{d})=\frac{s}{\sqrt{n-1}}=\frac{1.7058}{\sqrt{11}}=0.5143$$

$$t_c=\frac{\bar{d}}{SE(\bar{d})}=\frac{9.9167}{0.5143}=19.2819$$

$$t_c=19.2819$$

Step 6: Statistical decisions

Since $t_c = 19.2819 > 2.201$, according to the decision rule, we reject H_0.

Step 7: Conclusion

We conclude that, at 5% level of significance with 11 *df*, there is a significant difference in the mean production rate before and after change.

13.18 Chi-square Distribution

Chi-square distribution comes under the category of continuous probability distribution. It was first introduced by the Helmert (1875) and then remodified and introduced by Karl Pearson (1900).

The chi-square distribution can be mathematically defined as follows:

$$f(u) = \frac{1}{([v/2]-1)!2^{\frac{v}{2}}} u^{([v/2]-1)} \times e^{-u/2}; 0 < u < \infty$$

where $u = \sum_{i=1}^{n}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2$

and v is called the degrees freedom. The x_i are normally and independently distributed with mean, μ_i, and SD, σ_i.

It is denoted by the Greek letter, χ^2 (chi-squared).

Let X be a normally distributed random variable with mean, μ, and SD, σ.

Let us draw many independent random samples of size n from this population. Convert each value of x_i within each sample to the equivalent standard normal value. We have $z = \frac{x_i - \mu}{\sigma}; i = 1, 2, ...n$

Squaring and adding all the n items we have,

$$u = \sum_{i=1}^{n} z^2 = \sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2$$

We will have the sampling distribution of $U = \sum Z^2$, which is the chi-square $[\chi^2]$ distribution with n degrees of freedom.

The essentiality of chi-square distribution rests on the value for large samples.

$\chi^2 = \sum_{i=1}^{k}\left(\frac{O_i - E_i)^2}{E_i}\right)$ is distributed approximately as chi-square with v degrees of freedom.

NOTE: The degrees of freedom, v, can be evaluated differently for different situations. It will be explained at the time of evaluation.

where,

O_i = an observed frequency

E_i = an expected frequency

k = the number of pairs of observed and expected frequencies

NOTE:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{13.3}$$

$$= \sum_{i=1}^{k}\left(\frac{O_i^2 + E_i^2 - 2O_iE_i}{E_i}\right) = \sum_{i=1}^{k}\left(\frac{O_i^2}{E_i} + E_i^2 - 2O_i\right)$$

$$\chi^2 = \sum_{i=1}^{k} \frac{O_i^2}{E_1} + \sum_{i=1}^{k} E_i - 2\sum_{i=1}^{k} O_i$$

Since $\sum_{i=1}^{k} E_i = \sum_{i=1}^{k} O_i = m$, the total number of observations we have

$$\chi^2 = \sum_{i=1}^{k} \left[\frac{O_i^2}{E_i}\right] - m \tag{13.4}$$

Equations (13.3) and (13.4) are equivalent. For evaluation of chi-square calculated value, one can make use of either the formula stated in either Equation (13.3) or (13.4).

13.18.1 Properties of Chi-square Distribution

1. It is a continuous distribution with the probability density function is defined is $f(u) = \frac{1}{([v/2]-1)!2^{\frac{v}{2}}} u^{([v/2]-1)} \times e^{-u/2}$;
2. Its mean = v, SD = $\sqrt{2\nu}$, and mode = $v - 2$.
3. When $n > 1$, the probability curve is positively showed and starting from 0 extends to infinity on the right.
4. The sum of the two independent χ^2 – variants are also χ^2 variant.
5. $\sum_{i=1}^{k} \left(\frac{x_i - \mu}{\sigma}\right)^2$ follows chi-square distribution with k degrees of freedom.
6. $\sum_{i=1}^{k} \left(\frac{x_i - \bar{x}}{\sigma}\right)^2$ follows chi-square distribution with (k – 1) degrees of freedom.
7. It can be used for both large and small sample tests.
8. When $k \to \infty$, it turns to be a normal distribution.

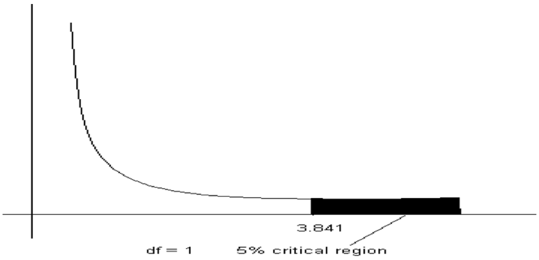
13.18.2 Chi-square Test

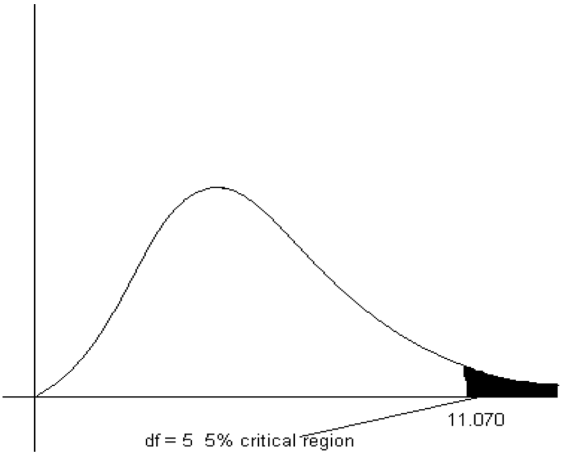
A chi-square test is widely used statistical test because of its simplicity. It can be used in the following three different situations:

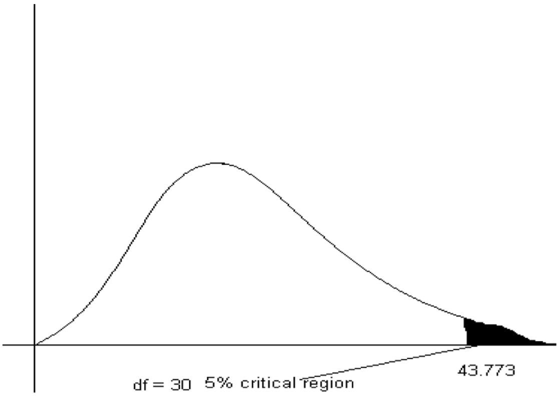
1. To test the goodness of fit
2. To test the independence of attributes
3. To test, whether the population has a specified value of the variance σ_0^2

NOTE: The chi-square distribution is a family of distributions and changes shape with changes in the number of degrees of freedom. For less degree of freedom, the distribution is badly skewed to the right. When the degrees of freedom are greater than or equal to 30, the distribution is approximately normal.

Critical value of chi-square with 5% level of significance with 1 *df*.





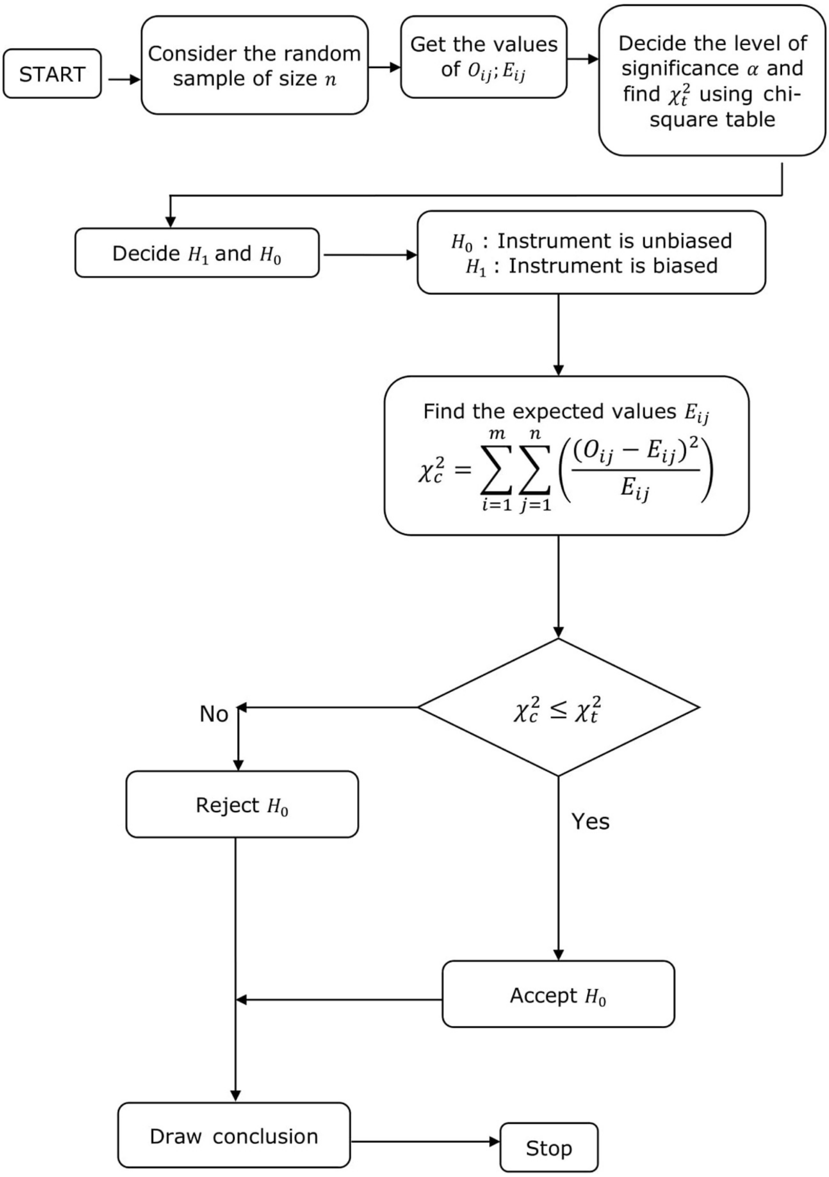


13.18.3 Test for Goodness of Fit

A test for goodness of fit is used to decide whether there is a significant difference between theory and experiment. Refer to Flowchart 13.9.

Example:

An ad agency collected information based on a questionnaire from a random sample, containing 60 viewers to indicate which of the 6 television programs he or she prefers. The results are as follows:


START
Consider the random sample of size n
Get the values of $O_{ij}; E_{ij}$
Decide the level of significance α and find χ_t^2 using chi-square table
Decide H_1 and H_0
H_0 : Instrument is unbiased
H_1 : Instrument is biased
Find the expected values E_{ij}
$\chi_c^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\right)$
$\chi_c^2 \le \chi_t^2$
No
Yes
Reject H_0
Accept H_0
Draw conclusion
Stop

FLOWCHART 13.9
Test for goodness of fit.

Program	1	2	3	4	5	6	Total
Number	5	8	10	12	12	13	60

Can we conclude from these data that the programs are not equally preferred?

Step 1: Here, $n = 6$, $v = n - 1 = 6 - 1 = 5$, the given data set is considered to be the observed frequencies and the same is notated by O_i ($i = 1, 2, \dots, 6$). With the concept of probability application, we have to evaluate the expected frequencies E_i ($i = 1, 2, \dots, 6$) for each observed frequency.

If we ascertain that the programs are equally probable, then

$$E_i = \left(\frac{1}{n}\right)(\text{total}) = \left(\frac{1}{6}\right)(60) = 10; \text{ for all } i = 1,2, \dots , 6.$$

Step 2: Framing the hypothesis

H_0: The 6 programs are equally preferred; H_1: The 6 programs are not equally preferred.

Step 3: Defining the test statistic

Because the study is related to the difference between the observed and expected frequencies, the test statistic to be evaluated is χ^2_c [chi-square calculated value]. It is defined as $\chi^2_c = \sum_{i=1}^{6}\left(\frac{(O_i-E_i)^2}{E_i}\right)$

Step 4: Defining the significance level

Because the level of significance α is not given, let us assume it as $\alpha = 0.05$. According to the table, the critical value of $\chi^2_t(0.05,5) = 11.07$. We see that if χ^2_c of the sample statistic is less than or equal to χ^2_t, then we are 95% confident that H_0 is true ($P(\chi^2_c \le 11.07) = 0.95$).

Step 5: Evaluate $E_i(i = 1, \dots , 6)$ and χ^2_c

Program	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$	O_i^2/E_i
1	5	10	−5	25	2.5	2.5
2	8	10	−2	4	0.4	6.4
3	10	10	0	0	0	10.0
4	12	10	2	4	0.4	14.4
5	12	10	2	4	0.4	14.4
6	13	10	3	9	0.9	16.9
Total					4.6	64.6

$$E_i = \left(\frac{1}{n}\right)[\text{total frequency}] = \frac{1}{6}(60) = 10$$

$$\chi^2_c = \sum_{i=1}^{6}\left(\frac{(O_i - E_i)^2}{E_i}\right) = 4.6$$

Alternative method: $\chi^2_c = \sum_{i=1}^{6}\left(\frac{O_i}{E_i}\right)^2 - m$; here $m = 60$.

$$\chi^2_c = 64.6 - 60 = 4.6$$

Step 6: Statistical decisions

Since $\chi^2_c = 4.6$, which is in the acceptance region ($4.6 \le 11.07$), according to the decision rule, we accept H_0.

Step 7: Conclusion

We conclude that at 5% level of significance with 5 *df*, all the 6 programs are equally preferred.

NOTE: Both the methods can be used interchangeably to evaluate the value of $\chi^2{}_c$. The second method is simpler than the first method.

Example:
The following table gives the number of aircraft accidents that occurred during various days of the week. Test whether the accidents are uniformly distributed over the week.

Days	Mon	Tue	Wed	Thur	Fri	Sat
Accidents	14	18	12	11	15	14

Given the values of test statistic significance at 5, 6, and 7, *df* are respectively 11.07, 12.5,9 and 14.07, at 5% level of significance.

Step 1: Here $n = 6$, $v = n - 1 = 6 - 1 = 5$; $m = 84$.

Given data set is considered to be the observed frequencies and the same is notated by $O_i(i = 1, 2, \dots, 6)$. With the concept of probability application, we have to evaluate the expected frequencies $E_i[i = 1, 2, \dots, 6]$ for each observed frequencies.

If we assume that the accidents are uniformly distribution over the week,

then $E_i = \frac{1}{n}$(total accidents) $= \frac{1}{6}(84) = 14$ for all $i = 1, 2, \dots, 6$.

Step 2: Framing the hypothesis

H$_0$: The accidents are uniformly distributed.

H$_1$: The accidents are not uniformly distributed.

Step 3: Defining the test statistic

Because the study is related to the difference between the observed and expected frequencies, the test statistic to be evaluated is $\chi^2{}_c$. It is defined as

$$\chi^2{}_c = \sum_{i=1}^{6}\left(\frac{(O_i - E_i)^2}{E_i}\right) = \sum_{i=1}^{6}\left(\frac{O_i}{E_i}\right)^2 - m$$

Step 4: Defining the significance level

The level of significance α is given as 0.05. According to the table, the critical value of $\chi^2{}_t\ (\alpha,\nu) = \chi^2{}_t\ (0.05,5) = 11.07$. We see that if $\chi^2{}_c$ of the sample statistic is less than or equal to $\chi^2{}_t$, then we are 95% confident that H$_0$ is true ($P(\chi^2{}_c \le 11.07) = 0.95$).

Step 5: Evaluate $E_i (i = 1, 2, \dots , 6)$ and χ^2_c

Days	O_i	E_i	O_i^2	O_i^2/E_i
Monday	14	14	196	14.00
Tuesday	18	14	324	23.14
Wednesday	12	14	144	10.29
Thursday	11	14	121	8.64
Friday	15	14	225	16.07
Saturday	14	14	196	14.00
Total	84			86.14

$$E_i = [1/n][\text{total frequency}] = [1/6][84] = 14; [i = 1, 2, \dots , 6]$$

$$\chi^2_c = 86.14 - 84 = 2.14.$$

Step 6: Statistical decisions

Since $\chi^2_c = 2.14$ is in the acceptance region ($2.14 \leq 11.07$), according to the decision rule, we accept H_0.

Step 7: Conclusion

We conclude that at 5% level of significance with 5 *df*, the number of accidents on different days are uniformly distributed.

Example:

A survey of 320 families with 5 children each revealed the following distribution:

No. of Boys	**5**	**4**	**3**	**2**	**1**	**0**
No. of Girls	0	1	2	3	4	5
No. of Families	14	56	110	88	40	12

Is the result consistent with the hypothesis that male and female births are equally probable?

Step 1:

Here, $n = 6$, $v = 6 - 1 = 5$, $m = 320$. Given data set is the observed frequencies and the same is notated by $O_i (i = 1,2, \dots , 5)$, with the concept of probability application, we have to evaluate the expected frequencies for each observed frequency.

If we assume that the male and female births are equally probable,

then, $E_i = 320\left({}^s c_i p^i q^{n-1}\right)$; $[i = 5,4,3,2,1,0]$

where $p = \frac{1}{2}; q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}; q = 0.5.$

$$E_i = 320\left(5_{c_1}\left(\frac{1}{2}\right)^i\left(\frac{1}{2}\right)^{5-i}\right); i = 5,4,3,2,1,0.$$

Step 2: Framing the hypothesis

H_0: The male and female births are equally probable.

H_1: The male and female births are not equally probable.

Step 3: Defining the test statistic

Because the study is related to the difference between the observed and expected frequencies, the test statistic to be evaluated is $\chi^2{}_c$. It is defined as

$$\chi^2{}_c = \sum_{i=1}^{6}\left(\frac{(O_i - E_i)^2}{Ei}\right) = \sum_{i=1}^{6}\left(\frac{O_i}{E_i}\right)^2 - m$$

Step 4: Defining the significance level

The level of significance α is not given, let us assume that $\alpha = 0.01$. According to the chi-square table, the critical value $\chi^2{}_t\ (\alpha, v) = \chi^2{}_t\ (0.01,5) = 15.09$. We see that if $\chi^2{}_c$ of the sample statistic is less than or equal to 15.09, then we are 99% confident that H_0 is true ($P(\chi^2{}_c \leq 15.01) = 0.99$).

Step 5: Evaluate $E_i(i = 5,4,3, \ldots , 0)$ and $\chi^2{}_c$.

Number of Male Children (*i*)	***O*$_i$**	***E*$_i$**	***O*$_i^2$**	***O*$_i^2$/*E*$_i$**
5	14	10	196	19.60
4	56	50	3136	62.72
3	110	100	12100	121.00
2	88	100	7744	77.44
1	40	50	1600	32.00
0	12	10	144	14.40
Total	320	320		327.16

where, $E_i = 320\left[{}^5C_i\left(\frac{1}{2}\right)^{5-i}\left(\frac{1}{2}\right)^i\right]$; $i = 5,4,3,2,1,0$.

When $i = 5$; $E_1 = 320 \times \left[{}^5C_5\left(\frac{1}{2}\right)^{5-5}\left(\frac{1}{2}\right)^5\right] = 320 \times 0.5^5 = 10$; similarly, one can evaluate all the other expected values.

$$\chi^2{}_c = \sum_{i=1}^{6}\left(\frac{O_i}{E_i}\right)^2 - m = 327.16 - 320 = 7.16.$$

Step 6: Statistical decisions

Since $\chi^2{}_c = 7.16$ lies in the acceptance region ($7.16 \leq 15.09$), according to the decision rule, we accept H_0.

Step 7: Conclusion

We conclude that at 1% level of significance with 5 *df*, male and female births are equally probable.

Example:

Sample analyses of examination results of 500 students were made. It was found that 220 students had failed, 170 had secured third class, 90 were placed in second class, and 20 got first class. Are these figures commensurate with general examination result, which is in the ratio 4:3:2:1 for the various categories, respectively?

[Table value of χ^2[3, 0.05] = 7.82]

Step 1:

Here $n = 4$, $v = 4 - 1 = 3$, $m = 500$

Given data set is the observed frequencies and the same is notated by O_i ($i = 1, \ldots, 4$), with the concept of probability application, we have to evaluate the expected frequencies for each observed frequency.

Given general examination result ratio:

Fail; third class; second class; first class as 4:3:2:1.

Step 2: Framing the hypothesis

H$_0$: The actual examination results and the general exam results are independent.

H$_1$: The actual examination results and the general exam results are dependent.

Step 3: Defining the test statistic

Because the study is related to the difference between observed and expected frequencies, the test statistic to be evaluated is x_c^2.

It is defined as $\chi^2{}_c = \left(\sum_{i=1}^{4}\left(\frac{O_i^2}{E_i}\right)\right) - m$

Step 4: Defining the significance level

The level of significance α is given as $\alpha = 0.05$. According to the chi-square table, the critical value is $x_t^2(\alpha, \nu) = x_t^2(0.05, 3) = 7.82$. We see that, if x_c^2 of the sample statistic is less than or equal to 7.82, then we are 95% confident that H$_0$ is true (P ($x_c^2 \le 7.81$) = 0.95).

Step 5: Evaluate E_i($i = 1, \ldots ,4$) and x_c^2

Examination Ranking	O_i	E_i	O_i^2/E_i
Fail	220	200	242.00
Third class	170	150	192.67
Second class	90	100	81.00
First class	20	50	8.00
Total	500	500	523.67

$$E_1 = \frac{500}{10}(4) = 200$$

$$E_2 = \frac{500}{10}(3) = 150$$

$$E_3 = \frac{500}{10}(2) = 100$$

$$E_4 = \frac{500}{10}(1) = 50.$$

$$\chi^2{}_c = \left(\sum_{i=1}^{4}\left(\frac{O_i^2}{E_i}\right)\right) - m = 523.67 - 500$$

$$\chi^2{}_c = 23.67$$

Step 6: Statistical decisions

Since $\chi^2{}_c = 23.67$ lies in the critical region ($23.67 > 7.82$), according to the decision rule, we reject H_0.

Step 7: Conclusion

We conclude that at 5% level of significance with 3 *df*, the actual examination results and the general examination results are dependent.

13.18.4 Tests for Independence of Attributes

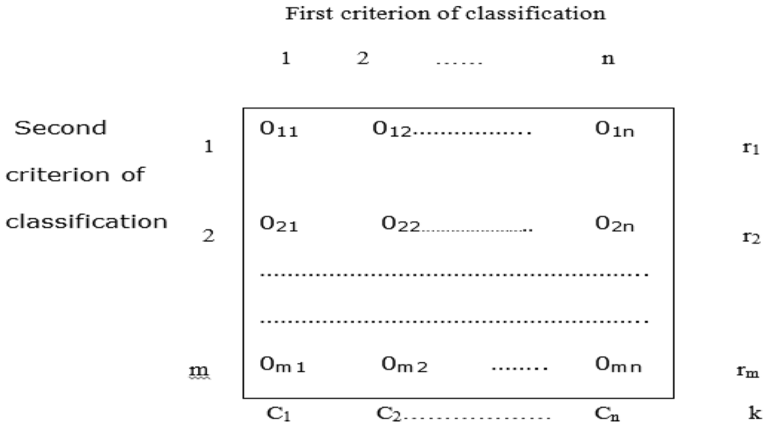
One of the most frequent uses of chi-square is for testing the H_0 that two criteria of classification, when applied to a population of subjects, are independent. It is said to be independent if the distribution of one criterion in no way depends on the distribution of another. If they are not independent, there is an association between them. (Refer to Flowchart 13.10.)

Contingency Table:

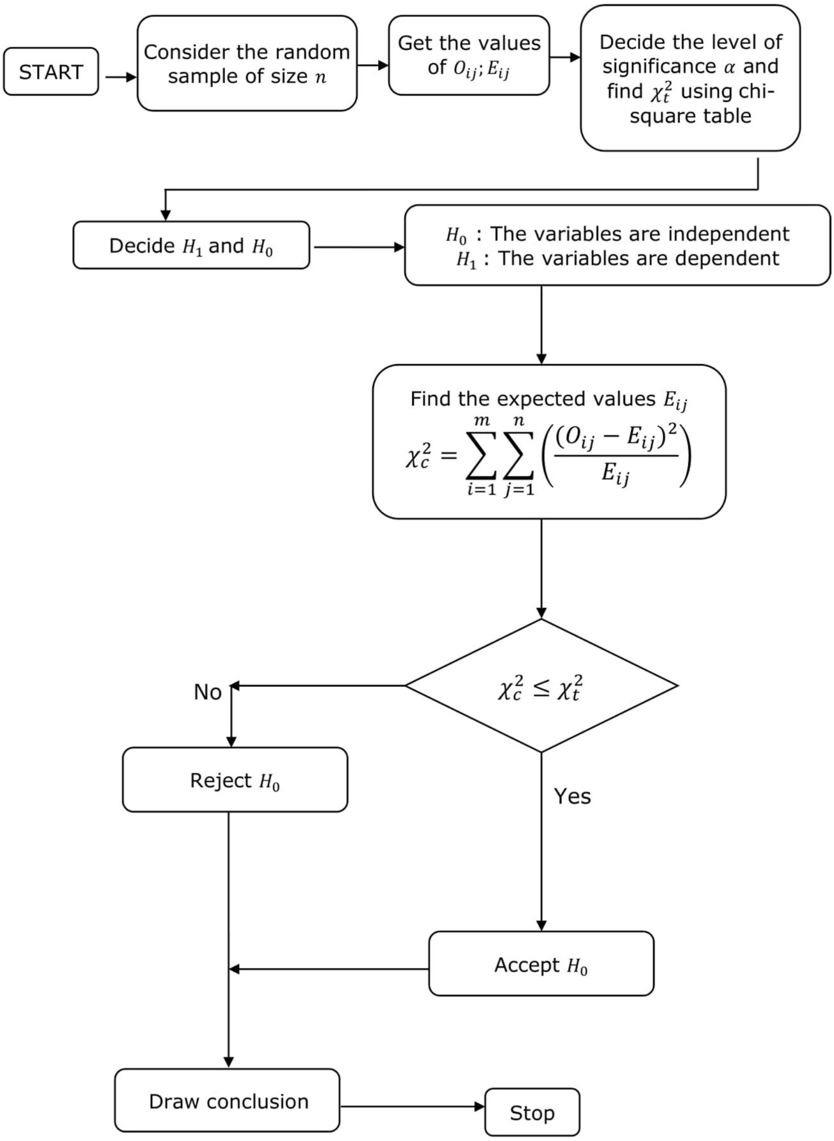
$\sum_{i=1}^{m} r_i = \sum_{j=i}^{n} c_j = k$. [It is obvious that the row summation and the column must be the same.]

The cell entries are referred to as observed frequency. If the two criteria of classification are independent, a joint probability is equal to the probability of two corresponding marginal probabilities.

Under the hypothesis of independence, the expected frequencies can be evaluated using the following relation:



		First criterion of classification				
		1	2	......	n	
Second criterion of classification	1	O11	O12................		O1n	r1
	2	O21	O22......................		O2n	r2
		....	....	....	....	
		....	....	....	....	
	m	Om 1	Om 2	....	Om n	rm
		C1	C2.................		Cn	k


START
Consider the random sample of size n
Get the values of $O_{ij}; E_{ij}$
Decide the level of significance α and find χ_t^2 using chi-square table
Decide H_1 and H_0
H_0 : The variables are independent
H_1 : The variables are dependent
Find the expected values E_{ij}
$\chi_c^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\right)$
$\chi_c^2 \le \chi_t^2$
No
Yes
Reject H_0
Accept H_0
Draw conclusion
Stop

FLOWCHART 13.10
Test for independence of attributes.

$$E_{ij} = \left(\frac{r_i}{k}\right)\left(\frac{c_j}{k}\right) \times k;\ i = 1 \ldots m \text{ and } j = 1, 2, \ldots ,n.$$

It can be simplified as, $E_{ij} = \dfrac{r_i * c_j}{k}$; for all $i = 1,2,.....m$ and $j = 1,2,...n$

For example, $E_{11} = \frac{r_1 \times c_1}{k}$; $E_{12} = \frac{r_1 \times c_2}{k}$; likewise one can evaluate all the values.

In an alternative way, the same values can be evaluated as follows:

$$E_{ij} = \frac{RT_i \times CT_j}{GT}$$

Where,
RT_i - ith row total
CT_j - jth column total
GT - Grand Total

NOTE: The cross-classification table is referred to as contingency table.

The degrees of freedom can be calculated by using the relation

$$df = (\text{numbers of rows} - 1) \times (\text{number of columns} - 1)$$

Once we have computed the expected frequency for each cell, the chi-square value can be evaluated as,

$$\chi^2{}_c = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{(O_{ij} - E_{ij})^2}{E_{ij}}\right)$$

As per the regular process, the calculated value of chi-square should be compared with the chi-square table value based on the level of significance given or assumed. Then the necessary action can be taken with reference to H_0.

NOTE: Each square box is called 'cell'.

Example:
A market research organization wants to determine whether there is a relationship between the size of a tube of toothpaste that a shopper buys and the number of persons in the shopper's household.

		NPH	**NPH**	**NPH**	**NPH**	
		1–2	**3–4**	**5–6**	**7 or More**	**Total**
STB	Giant	23	116	78	43	260
STB	Large	54	25	16	11	106
STB	Small	31	68	39	8	146
	Total	108	209	133	62	512

NPH, number of persons in household; STB, size of tube bought.

At the level of significance $\alpha = 0.01$, is there any relationship?

Step 1: Number of rows = $r = 3$; Number of columns = $c = 4$; $K = 512$

$$v = df = (r - 1) \times (c - 1) = (3 - 1)(4 - 1) = 6$$

Given data set is the observed frequencies and the same is notated by $O_{ij}(i = 1{,}2{,}3$ and $j = 1{,}2{,}3{,}4)$.

Name the cells using the alphabet either row wise or column wise. Here it is named row wise.

		NPH	**NPH**	**NPH**	**NPH**	
		1–2	*3–4*	*5–6*	**7 or More**	**Total**
STB	Giant	23-A	116-B	78-C	43-D	260
STB	Large	54-E	25-F	16-G	11-H	106
STB	Small	31-I	68-J	39-K	8-L	146
	Total	108	209	133	62	512

NPH, number of persons in household; STB, size of tube bought.

With the concept of probability application, we must evaluate the expected frequencies for each cell E_{ij} $(i = 1{,}2{,}3$, and $j = 1{,}2{,}3{,}4)$.

By definition, $E_{ij} = \frac{(r_i)(c_j)}{k}$; $i = 1{,}2{,}3$ and $j = 1{,}2{,}3{,}4$

Step 2: Framing the hypothesis

H$_0$: Total number of persons in the house and the size of the toothpaste purchased are independent.

H$_1$: Total number of persons in the house and the size of the toothpaste purchased are dependent.

Step 3: Defining the test statistic

Because the study is related to the independence of attributes, the test statistic to be evaluated is $\chi^2{}_c$.

It is defined as, $\chi^2{}_c = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\right)$

Step 4: Defining the level of significance

Because the level of significance α is given as 0.01, according to the chi-square table, the critical value of is $\chi^2{}_t(\alpha, \nu) = \chi^2{}_t(0.01, 6) = 16.81$.

We see that, if $\chi^2{}_c$ of the sample statistic is less than or equal to 16.81, then we are 99% confident that H$_0$ is true.

Step 5: Evaluate the values of E_{ij} $(i = 1{,}2{,}3;\ j = 1{,}2{,}3{,}4)$ and $\chi^2{}_c$

we can name each cell by using the alphabets.

$$E_{ij} = \frac{(r_i)(c_j)}{k} = \frac{(RT)_i \times (CT)_j}{(GT)} \text{ for all } i = 1{,}2{,}3;\ j = 1{,}2{,}3{,}4$$

$$\chi^2{}_c = \sum_{i=1}^{3}\sum_{j=1}^{4}\frac{(O_{ij}-E_{ij})^2}{(E_{ij})} = 89.34$$

Name of Cells	O_{ij}	$(RT)_i$	$(CT)_j$	E_{ij}	$O_{ij}–e_{ij}$	$(O_{ij}–E_{ij})^2$	$\frac{(O_{ij}-E_{ij})^2}{(E_{ij})}$	$\frac{O_{ij}^2}{E_{ij}}$
A	23	260	108	54.85	–31.85	1014.42	18.49	9.65
B	116	260	209	106.13	9.87	97.42	0.92	126.79
C	78	260	133	67.54	10.46	109.41	1.62	90.08
D	43	260	62	31.48	11.52	132.71	4.22	58.74
E	54	106	108	22.36	31.64	1001.09	44.77	130.41
F	25	106	209	43.26	2.85	8.12	7.71	14.45
G	16	106	133	27.54	−11.54	133.17	4.84	9.3
H	11	106	62	12.84	−1.84	3.39	0.26	9.42
I	31	146	108	30.8	0.2	0.04	0	31.20
J	68	146	209	59.6	8.4	70.56	1.18	77.58
K	39	146	133	37.93	1.07	1.14	0.03	40.10
L	8	146	62	17.68	−9.68	93.70	5.3	3.62
Total							89.34	601.34

Alternative method:

$$\chi^2{}_c = \left(\sum_{i=1}^{3}\sum_{j=1}^{4}\frac{O_{ij}{}^2}{E_{ij}}\right) - m = 601.34 - 512 = 89.34$$

NOTE: While comparing the calculation involved in evaluating the value of χ_c^2, it is better to use the alternative method.

Step 6: Statistical decisions

Since $\chi^2{}_c$= 89.34, is in the rejection region (89.34 > 16.81), according to the decision rule, we reject H_0.

Step 7: Conclusion

We conclude that at 1% level of significance with 6 *df*, the size of the toothpaste purchased depends on the total number persons in the house.

Example:
Two sample polls of votes for two candidates, A and B, for a public office are taken, one from among residents of rural areas and one from urban areas. The results are given in the following table. Examine whether the nature of the area is related to the voting preference in this election.

Votes for Area	A	B	Total
Rural	620-D	380-E	1000
Urban	550-F	450-H	1000
Total	1170	830	2000

[Given the tabulated value of test statistic with 1 degree of freedom and 5% level of significance is 3.841.]

Step 1: Number of rows = $r = 2$

Number of columns = $c = 2$

$k = 2000$ (Total sample size)

$$v = df = (r - 1) \times (c - 1) = (2 - 1) \times (2 - 1) = 1$$

Given data set is the observed frequencies and the same is notated by

$O_{ij}(i = 1,2,$ and $j = 1,2)$.

With the concept of probability application, we must evaluate the expected frequencies

$E_{ij}(i = 1, 2$ and $j = 1,2)$ for each cell.

By definition,

$$E_{ij} = \frac{(r_i)(c_j)}{k} = \frac{(RT)_i(CT)_j}{GT}; \; i = 1,2, \text{ and } j = 1,2.$$

Step 2: Framing the hypothesis

H$_0$: The nature of residence area and their voting preferences are independent.

H$_1$: The nature of residence area and their voting preference are dependent.

Step 3: Defining the test statistic

Because the study is related to the independence of attributes, the test statistic to be evaluated is $\chi^2{}_c$.

It is defined as

$$\chi^2{}_c = \sum_{i=1}^{2}\sum_{j=1}^{2}\left(\frac{(O_{ij} - E_{ij})^2}{E_{ij}}\right) = \left[\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{O_{ij}{}^2}{E_{ij}}\right] - k$$

Step 4: Defining the significance level

The level of significance α is given as 0.05. According to the chi-square table, the critical value is $\chi^2{}_c(\alpha, v) = \chi_t^2(0.05,1) = 3.841$. We see that, if $\chi^2{}_c$ of the sample statistic is less than or equal to 3.841, then we are 95% confident that H$_0$ is true.

Step 5: Evaluate $E_{ij}(i = 1,2$ and $j = 1,2)$ and $\chi^2{}_c$

Before that, we can name the cells by using alphabets.

$$E_{ij} = \frac{(r_i)(c_j)}{k} = \frac{(RT)_i(CT)_j}{GT}; \text{ for all } i = 1,2 \text{ and } j = 1,2$$

Name of the Cells	O_{ij}	$(RT)_i$	$(CT)_j$	E_{ij}	O_{ij}^2	O_{ij}^2/E_i
D	620	1000	1170	585	384400	357.09
E	380	1000	830	415	144400	347.95
F	550	1000	1170	585	302500	517.09
H	450	1000	830	415	202500	487.95
Total	2000			2000		2010.08

$$\chi^2_c = \left(\sum_{i=1}^{2} \sum_{j=1}^{2} \frac{O_{ij}^{\ 2}}{E_{ij}} \right) - k \ \chi^2_c = 2010.08 - 2000 = 10.08$$

$$\chi^2_c = 10.08$$

Step 6: Statistical decisions

Since $\chi^2_c = 10.08$ is in the critical region (10.08 > 3.841), according to the decision rule, we reject H_0.

Step 7: Conclusion

We conclude that at 5% level of significance with 1 *df*, the nature of area and their voting preferences are dependent.

Example:

Two researchers adopted two different sampling techniques while investigating the same group of students to find the number of students falling in to different intelligence levels. The results are as follows:

Step 1: Number of rows = $r = 2$

Number of columns = $c = 4$.

$$k = 300;\ v = df = (r - 1)(c - 1) = (2 - 1)\ (4 - 1) = 3$$

Research/No. of Students	<Average	Average	>Average	Excellent	Total
1	86	60	44	10	200
2	40	33	25	02	100
Total	126	93	69	12	300

Would you say that the sampling techniques adopted by the two researchers are independent?

Given data set is the observed frequencies and the same is notated by

$O_{ij}(i = 1{,}2$, and $j = 1{,}2{,}3{,}4)$. With the concept of probability application, we have to evaluate the expected frequencies $E_{ij}(i = 1, 2$ and $j = 1, \ldots, 4)$ for each cell.

By definition, $E_{ij} = \frac{(r_i)(c_j)}{k} = \frac{(RT)_i(CT)_j}{GT}$; for all $i = 1{,}2$ and $j = 1{,}2{,}3{,}4$.

Step 2: Framing the hypothesis

H_0: The sampling technique adopted by the researchers is independent.

H_1: The sampling technique adopted by the researchers is dependent.

Step 3: Defining the test statistic

Because the study is related to the independence of attributes, the test statistic to be evaluated is χ^2_c.

It is defined as, $\chi^2_c = \left(\sum_{i=1}^{2}\sum_{j=1}^{4}\frac{[O_{ij}-E_{ij}]^2}{E_{ij}}\right)$

Step 4: Defining the level of significance

Because the level of significance α is not given, let us assume that $\alpha = 0.05$. According to the chi-square table, the critical value is

$$\chi_t^2(\alpha, v) = \chi_t^2(0.05, 2^*) = 5.001$$

$$\chi_t^2(\alpha, v) = \chi_t^2(0.09, 3^*) = 7.82$$

We see that, if χ^2_c of the sample statistic is less than are equal to the critical value, then we are 95% confident that H_0 is true.

Step 5: Evaluate $E_{ij}(i = 1,2$ and $j = 1,2,3,4)$ and χ^2_c.

Before that, we name the cells by using the alphabet.

$$E_{ij} = \frac{(r_i)(c_j)}{k} = \frac{(RT)_i(CT)_j}{GT} \text{ for all } i = 1,2, \text{ and } j = 1,2,3,4.$$

Name of the Cells	O_{ij}	$(RT)_i$	$(CT)_j$	E_{ij} ($GT = 300$)	O_{ij}^2	O_{ij}^2/E_{ij}
A	86	200	126	84	7396	88.05
B	60	200	93	62	3600	58.06
C	44	200	69	46	1936	42.09
D	10	200	12	8	100	12.50
E	40	100	126	42	1600	38.10
F	33	100	93	31	1089	35.13
G	25	100	69	23	729	27.00
H	2 } 27	100	12	4 } 27		
Total	300			300		300.93

Since the $E_{24} = 4$, according to the recommendations of Cochran, we can combine adjacent rows in such a way as to make the cell entry ≥5. According to this, we merge the two rows G and H.

NOTE: Due to this merging of two rows into a single row, the actual *df* should be reduced by 1. Current value of $v = 3$.

The modified $v = v - 1 = 3 - 1 = 2$. While comparing the critical value, we must consider χ^2_t (.05, 2) = 5.991.

$$\chi^2_c = \left(\sum_{i=1}^{2}\sum_{j=1}^{4}\left(\frac{O_{ij}^2}{E_{ij}}\right)\right) - k;\ \chi^2_c = 300.93 - 300.00 = 0.93.$$

Step 6: Statistical decisions

Since $\chi^2_c = 0.93$ is in the acceptance region $(0.93 \le 5.991)$, according to the decision rule, we accept H_0.

Step 7: Conclusion

We conclude that at the 5% level of significance with 2 *df*, there is no significance difference between the two sampling techniques adopted by researchers 1 and 2.

13.18.5 Whenever the Expected Frequencies of the Cell Entries Are Less Than 5

Case 1: In contingency–table analysis, some cells may have small expected frequencies (<5). This poses a possible threat to the validity of the chi-square test. At this critical situation, most statisticians overcome this type situation by following the recommendation given by Cochran. According to him, whenever the *df* of the related problem is more than 1, a minimum expected frequency per cell of 1 is permissible if no more than 20% of the cells have expected frequencies of less than 5. We may combine adjacent rows and columns to satisfy this rule, so long as this does not violate the logic of the classification scheme.

Case 2: If the degree of freedom is 1, we apply a correction as per F. Yates (1934) called the 'Yates correction', whenever any of the theoretical cell frequencies is less than 5.

Consider the 2 × 2 contingency table

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	$a+b+c+d$

Case 1: If all the cell entries are greater than or equal to 5, the χ^2_c value can be evaluated directly using the following formula:

$$\chi^2_c = \frac{k(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

Case 2: If any of the cell entries is less than 5, the χ^2_c value can be evaluated using the following Yates' correction formula:

$$\chi^2_c = \frac{k\ \left[\ |ad-bc| - \ [N/2]\right]^2}{(a+c)(b+d)(a+b)(c+d)}$$

NOTE: Whenever k is large, $|\{\chi^2_c\} - \{\chi^2_c \text{ (Yates correction)}\}|$ will be very small.

Example:

With the help of chi-square test, verify that whether the medicine is effective in preventing tuberculosis (TB).

	Affected by TB	Not Affected by TB	Total
Medicine group	31-A	469-B	500
Nonmedicine group	185-C	1315-D	1500
Total	216	1784	2000

Step 1: Number of rows = $r = 2$

Number of columns = $c = 2$

$$K = 2000$$

$$v = (r-1)(c-1) = (2-1)(2-1) = 1.$$

Given data set is the observed frequencies and the same is notated by $O_{ij}(i = 1,2$ and $j = 1,2)$ and with the concept of probability application, we must evaluate the expected frequencies $E_{ij}(i = 1,2$ and $j = 1,2)$ for each cell. Name each cell by using alphabet.

By definition

$$E_{ij} = \frac{(r_i)(c_j)}{k} = \frac{(RT)_i \times (CT)_j}{GT} \text{ for all } i = 1,2, \text{ and } j = 1, 2.$$

Step 2: Framing the hypothesis

H$_0$: The medicine and TB are independent (medicine is not effective)

H$_1$: The medicine and TB are dependent (medicine is effective)

Step 3: Defining the test statistic

Because the study relates to the independence of attributes, the test statistic to be evaluated is χ^2_c.

It is defined as,

$$\chi^2_c = \sum_{i=1}^{2}\sum_{j=1}^{2}\left(\frac{O_{ij}{}^2}{Eij}\right) - k$$

Step 4: Defining the significance level

Because the level of significance α is not given, let us assume that $\alpha = 0.05$. According to the chi-square table, the value of $\chi^2_t\ (\alpha, v) = \chi^2_t(0.05, 1) = 3.841$.

We see that, if χ^2_c of the sample statistic is less than or equal to 3.841, then we are 95% confident that H$_0$ is true.

Step 5 Evaluate $E_{ij}(i = 42$ & $j = 1,2)$ and χ^2_c.

Name of the Cells	O_{ij}	$(RT)_i$	$(CT)_j$	E_{ij} ($GT = 2000$)	$O_{ij}{}^2$	$O_{ij}{}^2/E_{ij}$
A	31	500	216	54	961	17.796
B	469	500	1784	446	219961	493.186
C	185	1500	216	162	34225	211.265
D	1315	1500	1784	1338	1729225	1292.395
Total	2000					2014.642

$$\chi^2_c = \left(\sum_{i=1}^{2}\sum_{j=1}^{2}\left(\frac{O_{ij}{}^2}{E_{ij}}\right)\right) - k = 2014.66 - 2000 = 14.642 = 14.64$$

Alternative method:

31	469	$(a + b)$
(a)	(b)	500
185	1315	$(c + d)$
(c)	(d)	1500
$(a + c)$	$(b + d)$	$(a + b + c + d)$
216	1784	2000

$$\chi^2_c = \frac{k(ad - bc)^2}{(a+b)(a+c)(c+d)(b+d)}$$

$$\chi^2_c = \frac{2000 \times [(31 \times 1315) - (185 \times 469)]^2}{500 \times 216 \times 1500 \times 1784} = 14.64;\ \chi^2_c = 14.64$$

Step 6: Statistical decisions

Since $\chi^2_c = 14.64$ is in the rejection region (14.64 > 3.841), according to the decision rule, we reject H_0.

Step 7: Conclusion

We conclude that at the 5% level of significance with 1 *df*, the medicine is quite effective with respect to the TB.

Example:

In an experiment on the immunization of goats from anthrax, the following results were obtained. Derive your inference on the variance.

	Diet of Anthrax	Survived	Total
Inoculated with vaccine	2	10	12
Not inoculated	6	6	12
Total	8	16	24

Step 1: Number of rows = $r = 2$

Number of Columns = $c = 2$

$$k = 24$$

$$v = (r - 1) \times (c - 1) = (2 - 1)(2 - 1) = 1$$

Given data set is the observed frequencies and the same is notated by $O_{ij}(i = 1,2$ and $j = 1,2)$ probability.

Note that the O_{11} value is less than 5 and $df = 1$, hence, for evaluation of χ^2_c, we can make use of Yates correction formula directly (no need to compare $E_{ij}(i = 1, 2$ and $j = 1,2)$.

Step 2: Framing the hypothesis

H_0: There is no relationship between the vaccine and the anthrax disease.

H_1: There is a relationship between the vaccine and the anthrax disease.

Step 3: Defining the test statistic

Because the study is related to the independence of attributes, the test statistic to be evaluated is χ^2_c. It is defined as,

$$\chi_c^2 = \frac{k\ \left[|\ \mathrm{a}d - bc| - \left(k/2\right)\ \right]^2}{(a+c)(b+d)(a+b)(c+d)}$$

Step 4: Defining the significance level

Because the level of significance is not given, let us assume that $\alpha = 0.05$. According to the chi-square table the value of

$$\chi_t^2(\alpha, v) = \chi_t^2(0.05, 1) = 3.841.$$

We see that, if χ^2_c of the sample statistic is less than or equal to 3.841, then we are 95% confident that H_0 is true.

Step 5: Evaluate χ^2_c using Yate's correction formula.

$$\chi_c^2 = \frac{24 \times (|12 - 60| - (24/2))^2}{12 \times 12 \times 8 \times 16} = 1.6875;\ \chi^2_c = 1.6875$$

Step 6: Statistical decisions

Since $\chi^2_c = 1.6845$ is in the acceptance region $(1.6845 \leq 3.841)$, according to the decision rule, we accept H_0.

Step 7: Conclusion

We conclude that at the 5% level of significance with 1 *df*, the disease and the medicine are independent.

NOTE:

$$\chi_c^2 = \frac{(n-1)s^2}{\sigma_0^2};\ s^2 = \frac{\sum_{i=1}^{n}(x_1 - \overline{x})^2}{(n-1)}$$

13.18.6 Test for a Specified Population Variance

Consider a random sample of n-item; $X_1, X_2, \ldots, X_n$ out of a normal population with mean $\overline{x}$ and variance s^2.

The test is to verify whether the population variance can be equal to the specified value of variance σ_o^2. Then, H_0: $\sigma^2 = \sigma_o^2$; H_1: $\sigma^2 \neq \sigma_o^2$

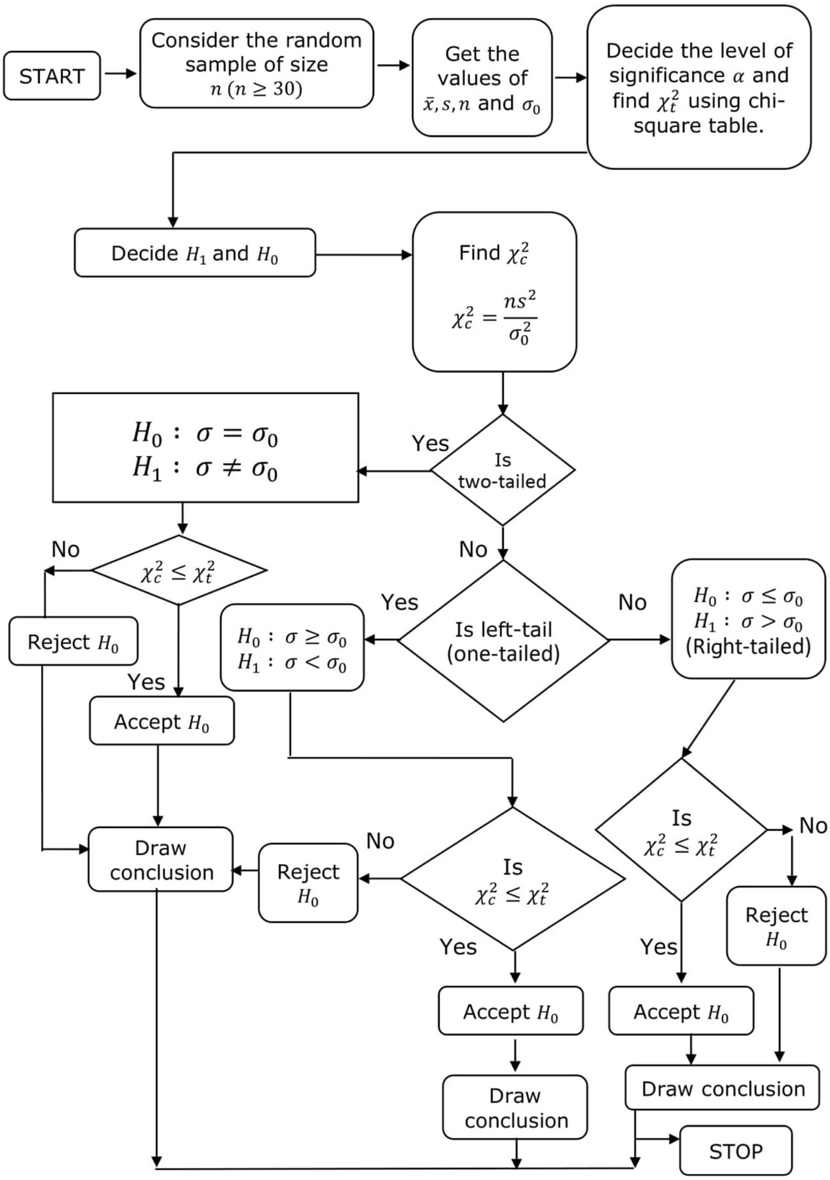
The test statistic is χ_c^2.

$$\chi_c^2 = \frac{ns^2}{\sigma_o^2};\ \text{where } s^2 = \sum_{i=1}^{n}\frac{(x_1 - \overline{x})^2}{n} \text{ or } \chi_c^2 = \sum_{i=1}^{n}\frac{(x_1 - \overline{x})^2}{\sigma_o^2}$$

Compare the $\chi_c{}^2(\alpha, v)$ with the $\chi_t{}^2(\alpha, v)$. If $\chi_c{}^2 \leq \chi_t{}^2$ accept H_0; if not reject H_0. (Refer to Flowchart 13.11.)

Example:

Consider the weight of the 10 different students in kilograms 49, 52, 48, 55, 43, 47, 53, 45, 40, and 38. Can we say the variance of the population is in which the sample is drawn is 20?


START
Consider the random sample of size $n\ (n \geq 30)$
Get the values of $\bar{x}, s, n$ and σ_0
Decide the level of significance α and find χ_t^2 using chi-square table.
Decide H_1 and H_0
Find χ_c^2
$\chi_c^2 = \frac{ns^2}{\sigma_0^2}$
Is two-tailed
Yes
$H_0 : \sigma = \sigma_0$
$H_1 : \sigma \neq \sigma_0$
No
$\chi_c^2 \leq \chi_t^2$
No
Reject H_0
Yes
Accept H_0
Draw conclusion
Is left-tail (one-tailed)
Yes
$H_0 : \sigma \geq \sigma_0$
$H_1 : \sigma < \sigma_0$
No
$H_0 : \sigma \leq \sigma_0$
$H_1 : \sigma > \sigma_0$
(Right-tailed)
Is $\chi_c^2 \leq \chi_t^2$
No
Reject H_0
Yes
Accept H_0
Draw conclusion
Is $\chi_c^2 \leq \chi_t^2$
No
Reject H_0
Yes
Accept H_0
Draw conclusion
STOP

FLOWCHART 13.11
Test for a specified variance.

Step 1: Consider the given data and find $\bar{x}, s$. Here n = 10; v = 10 – 1 = 9, σ_o = 20

x	$\bar{x} = 47; x - \bar{x}$	$(x - \bar{x})^2$
49	2	4
52	5	25
48	1	1
55	8	64
43	–4	16
47	0	0
53	6	36
45	–2	4
40	–7	49
38	–9	81
470		280

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{n} = \frac{470}{10} = 47$$

$$s^2 = \frac{\sum_{i=1}^{10} (x_i - \bar{x})^2}{n} = \frac{280}{10} = 28$$

$$\text{Hence, } \bar{x} = 47;\ s^2 = 28;\ n = 10;\ v = 9$$

Step 2: Framing the hypothesis

H_0: $\sigma^2 = 20$

H_1: $\sigma^2 \neq 20$

Step 3: Defining the test statistic

Because the study is related to the specified population variance, the test statistic to be evaluated is χ_c^2. It is defined as,

$$\chi_c^2 = \frac{\sum_{i=1}^{n} (x_1 - \bar{x})^2}{\sigma_o^2} = \frac{ns^2}{\sigma_o^2}$$

Step 4: Defining the significance level

Because the level of significance α is not given, let us assume that α = 0.05. According to the chi-square table, $\chi_t^2(\alpha, v) = \chi_t^2(0.05, 9) = 16.92$.

We see that if χ_c^2 of the sample statistic satisfies the condition $\chi_c^2 \leq 16.92$, then we are 95% confident that H_0 is true ($P(\chi_c^2 \leq 16.92) = 0.95$).

Step 5: Evaluate x_c^2

$$\chi_c^2 = \frac{ns^2}{\sigma_o^2} = \frac{10(28)}{20} = 14$$

Step 6: Statistical decisions

Since $x_c^2 = 14$, lies in the acceptance region (14 ≤ 16.92), according to the decision rule, we accept H_0.

Step 7: Conclusion

We conclude that at the 5% level of significance with 9 *df*, the population variance can be 20 kg^2.

NOTE: If sample variance is evaluated using the relation $S^2 = \frac{\sum_{i=1}^{10}(x_1-\bar{x})^2}{n-1}$; then the value of $\chi_c^2 = \frac{(n-1)S^2}{\sigma_0^2}$.

Example:

The tensile strength of a synthetic fibre must have a variance of 5 or less before it is acceptable to a certain manufacturer. A random sample of 25 specimens taken from a new shipment gives a variance of 7. Does this provide enough grounds for the manufacturer to refuse the shipment? Let α = 0.05, and assume that tensile strength of the fibre is proximately normally distributed.

Step 1: Consider the given data $\sigma_o^2 = 5$

$$n = 25, S^2 = \frac{\sum(x_1 - \overline{x})^2}{n-1} = 7$$

$$v = 25 - 1 = 24.$$

Step 2: Framing the hypothesis

$$H_0 : \sigma^2 \leq 5$$

$$H_1 : \sigma^2 > 5$$

Step 3: Defining the test statistic

Because the study is related to the specified population variance, the test statistic the evaluated is X^2_c.

It is defined as $\chi_c^2 = \frac{[n-1]s^2}{\sigma_o^2}$.

Step 4: Defining the significance level

Because the level of significance α is given as 0.05, according to the chi-square table, $\chi_t^2(\alpha, v) = \chi_t^2(005, 24) = 36.415$.

We see that if χ_c^2 of the sample statistic satisfies the condition $\chi_c^2 \leq 36.415$, then we are 95% confident that H_0 is true ($P(\chi_c^2 \leq 36.415) = 0.95$).

Step 5: Evaluate the value of χ_c^2

$$\chi_c^2 = \frac{[n-1]S^2}{\sigma_o^{\ 2}} = \frac{[24]*7}{5} = 33.6$$

$$\chi_c^2 = 33.6$$

Step 6: Statistical decisions

Because $\chi_c^2 = 33.6$ lies in the acceptance region ($33.6 \leq 36.415$), according to the decision rule, we accept H_0.

Step 7: Conclusion

We conclude that at the 5% level of significance with 24 *df*, the manufacturer should accept the shipment.

13.19 Snedecor's F-Distribution

This type of distribution falls under continuous probability distribution type introduced by G. W. Snedecor. To honour R. A. Fisher this was named F-distribution. Decisions about the equality of two population variances are based on the F-test. It is otherwise called a 'variance ration test'.

Let (S_1, n_1) and (S_2, n_2) be the two independent random samples. $S_1 = \{x_1, x_2......., x_{n1}\}$ and $S_2 = \{y_1, y_2......., y_{n2}\}$. Then their means and variances can be defined as,

$$\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_i}{n_1}; \bar{x}_2 = \frac{\sum_{i=1}^{n_2} y_i}{n_2}$$

$$S_1^{\ 2} = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2}{n_1 - 1}; S_2^{\ 2} = \frac{\sum_{i=1}^{n_2} (y_i - \bar{x}_2)^2}{n_2 - 1}$$

Then we define the statistic F by the relation $F = \frac{(S^2{}_1/\sigma_1{}^2)}{(S^2{}_2/\sigma_2{}^2)}$; since we assume that $\sigma_1{}^2 = \sigma_2{}^2$, we have $F_c = (S^2{}_1 / S_2{}^2); if\ S_1{}^2 > S_2{}^2$ [or] $(S^2{}_2 / S_1{}^2)$, if $S_2{}^2 > S_1{}^2$

Nature of Various	*F* Statistic	Numerate Degree of Freedom	Denominator Degrees of Freedom	Notation for Table Value's (α-level of Significance.)
$S_1^{\ 2} > S_2^{\ 2}$	$F_c = \frac{S_1^2}{S_2^2}$	$n_1 - 1 = v_1$	$n_2 - 1 = v_2$	$F_t(v_{1,} v_{2,}\ \alpha)$
$S_2^{\ 2} > S_2^{\ 2}$	$F_c = \frac{S_2^2}{S_1^2}$	$n_1 - 1 = v_1$	$n_2\ 1 = v_2$	$F_t(v_{2,} v_{1,}\ \alpha)$

The probability density function of F can be defined as,

$P(f) = k \times F^{([v_1/\ 2]-1)} \times \left(1+\frac{v_1F}{v_2}\right)^{-\left(\frac{v_1+v_2}{2}\right)}$; $0 \le F < \infty$. Where v_1 and v_2 are the *df* of the two estimates. *k* refers constant, and it can be evaluated using the relation $\int_0^{\infty} P[F]dF = 1$.

13.19.1 Properties of F-Distribution

1. The distribution *F* depends only on the two degrees of freedom v_1 and v_2.
2. It is positively skewed and starts from 0, rises to the peak at the value equal to $\frac{n_2(n_1-2)}{n_1(n_2+2)}$, and then falls to 0 as F increases without limit.
3. Its mean $\bar{x} = \frac{v_2}{v_2-2}$ and
4. variance $= \left(\frac{v_2}{v_2-2}\right)^2 \left(\frac{2(v_1+v_2-2)}{v_1(v_2-4)}\right) = \frac{2v_2^{\,2}(v_1+v_2-2)}{v_1(v_2-4)(v_2-2)^2}$; it is obvious to see that, $\bar{x}$ can be computed if $u_2 > 2$ and variance can be evaluated if $v_2 > 4$.

Selecting the appropriate test statistic for *F*.

Nature of Test	H_0	H_1	Appropriate Test Statistic
Two-sided test	$\sigma_1^{\,2} = \sigma_2^{\,2}$	$\sigma_1^{\,2} \neq \sigma_2^{\,2}$	$F = \frac{S_1^{\,2}}{S_2^{\,2}}$ or $\frac{S_2^{\,2}}{S_1^{\,2}}$ $(S_1 > S_2)$ or $(S_1 < S_2)$
One-sided test (right-tailed)	$\sigma_1^{\,2} \le \sigma_2^{\,2}$	$\sigma_1^{\,2} > \sigma_2^{\,2}$	$F = \frac{s_1^2}{s_2^2}$
One-sided test (left-tailed)	$\sigma_1^{\,2} \ge \sigma_2^{\,2}$	$\sigma_1^{\,2} < \sigma_2^{\,2}$	$F = \frac{s_2^2}{s_1^2}$

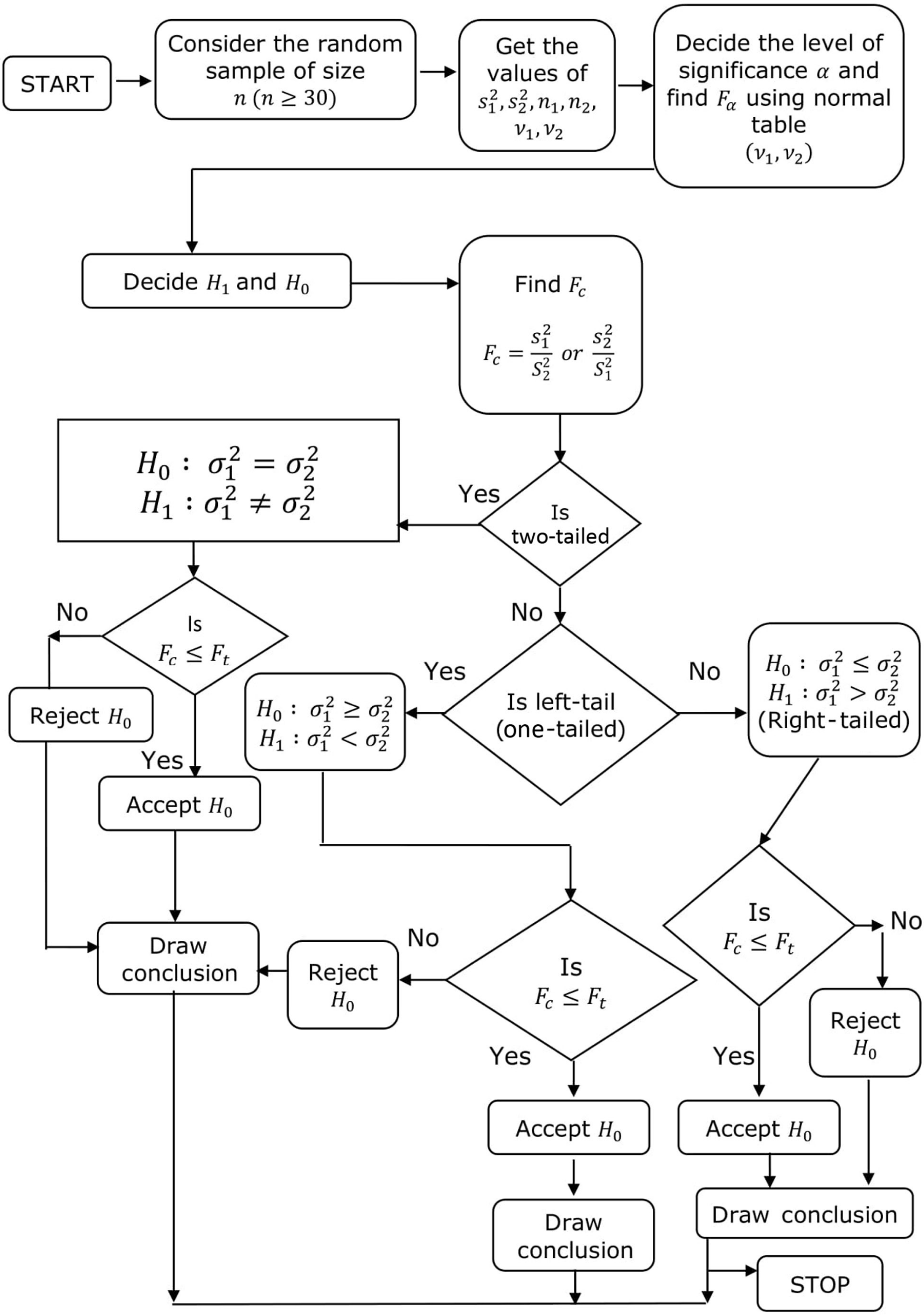
13.19.2 Test for Difference of Two Populations' Variances

Refer to Flowchart 13.12.

Example:

A firm wants to make a choice amongst certain makes of cycle tubes. It gathered information on the average running life and bursting strength of tubes based on samples drawn at random from large lots of those makes. The information about the two makes A and B is given here. The firm wants to know if the variances of the two makes are significantly different by applying F-test at the 5% level of significance.

	Brand A	Brand B
Sample size	21	16
SD	2.5	1.5
Mean running life	100	95


START
Consider the random sample of size $n\ (n \geq 30)$
Get the values of $s_1^2, s_2^2, n_1, n_2, \nu_1, \nu_2$
Decide the level of significance α and find F_α using normal table (ν_1, ν_2)
Decide H_1 and H_0
Find F_c
$F_c = \frac{s_1^2}{S_2^2}\ or\ \frac{s_2^2}{S_1^2}$
Is two-tailed
Yes
$H_0: \sigma_1^2 = \sigma_2^2$
$H_1: \sigma_1^2 \neq \sigma_2^2$
No
Is $F_c \leq F_t$
No
Reject H_0
Yes
Accept H_0
Draw conclusion
Is left-tail (one-tailed)
Yes
$H_0: \sigma_1^2 \geq \sigma_2^2$
$H_1: \sigma_1^2 < \sigma_2^2$
No
$H_0: \sigma_1^2 \leq \sigma_2^2$
$H_1: \sigma_1^2 > \sigma_2^2$
(Right-tailed)
Is $F_c \leq F_t$
No
Reject H_0
Draw conclusion
Yes
Accept H_0
Draw conclusion
Is $F_c \leq F_t$
No
Reject H_0
Yes
Accept H_0
Draw conclusion
STOP

FLOWCHART 13.12
Test for difference of two population variance.

Step 1:

Given that

	Sample 1	**Sample 2**
	Brand A	**Brand B**
Mean	$\overline{x}_1 = 100$	$\overline{x}_2 = 95$
SD	$s_1 = 2.5$	$s_2 = 1.5$
Sample size	$n_1 = 21$	$n_2 = 16$
df	$v_1 = n_1 - 1 = 20$	$v_2 = n_2 - 1 = 15$

Find S_1^2 and S_2^2, using the values of s_1^2 and s_2^2 using the relation.

$n_1 s_1^2 = (n_1 - 1)\, S_1^2$	$n_2 s_2^2 = (n_2 - 1)\, S_2^2$
$S_1^2 = \left(\frac{n_1}{n_1 - 1}\right) s_1^2$	$S_2^2 = \left(\frac{n_2}{n_2 - 1}\right) s_2^2$
$S_1^2 = \left(\frac{21}{20}\right)(2.5)^2$	$S_2^2 = \left(\frac{16}{15}\right)(1.5)^2$
$S_1^2 = 6.5625$	$S_2^2 = 2.4$

Let σ_1^2 and σ_2^2 stand for the two population variances.

Step 2: Framing the hypothesis

$$\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2$$

$$\mathrm{H}_1 : \sigma_1^2 \neq \sigma_2^2$$

Step 3: Defining the test statistic

Because the study is related to the difference of two population variances, the test statistic to be evaluated is F_c, and it is defined as,

$$F_c = \frac{s_1^2}{s_2^2}; \text{ since } S_1^2 > S_2^{2.}$$

Step 4: Defining the significance level

Because the level of significance α is not given, let us assume that $\alpha = 0.05$. Since we have a two-sided alternative (two-tailed test) [the value α to be consider as [$\alpha/2$]], according to the F-table, $F_t\left(\frac{\alpha}{2}, v_1, n_2\right) = F_t(0.025, 20, 15) = 2.7559$. We see that, if F_c of the sample statistic satisfies the condition $F_c \leq 2.755$, then we are 95% confident that H_0 is true ($P(F_c \leq 2.7559) = 0.95$).

NOTE: If $F = \frac{s_2^2}{s_1^2}(s_2^2 > s_1^2)$, the numerator degrees of freedom is v_2, and the denominator degrees of freedom is, $V_{1.}$

Step 5: Evaluate F_c

$$F_c = \frac{s_1^2}{s_2^2} = \frac{6.5625}{2.4} = 2.7344$$

$$F_c = 2.7344$$

Step 6: Statistical decisions

Since F_c = 2.7344 lies in the acceptance region (2.7344 < 2.7551), according to the decision rule, we accept H_0.

Step 7: Conclusion

We conclude that at the 5% level of significance with (20, 15) degrees of freedom, the difference between the two population variances is not significant.

Example:

Two sources of raw materials are under consideration by a company. Both sources seem to have similar characteristics, but the company is not sure about their respective uniformity. A sample of 10 lots from Source A yields a variance of 225 and a sample of 11 lots from Source B yields a variance of 200. Is it likely that the variance of Source A is significantly greater than the variance of Source B. $\alpha = 0.01$.

Step 1: Given that:

	Sample 1	**Sample 2**
	Source A	**Source B**
Variance	$s_1^2 = 225$	$s_2^2 = 200$
Size	$n_1 = 10$	$n_2 = 11$
df	$v_1 = 10 - 1 = 9$	$v_2 = 11 - 1 = 10.$

Find S_1^2 and S_1^2 using the value of s_1^2 and s_2^2 using the relation.

$$S_1^2 = \left(\frac{n_1}{n_1 - 1}\right)s_1^2\ ;\ S_2^2 = \left(\frac{n_2}{n_2 - 1}\right)s_2^2$$

$$S_1^2 = \left(\frac{10}{9}\right)(225) = 250$$

$$S_2^2 = \left(\frac{11}{10}\right)(240) = 264$$

Let σ_1^2 and σ_2^2 be the two population variances.

Step 2: Framing the hypothesis

$$H_0 : \sigma_1^2 \le \sigma_2^2$$

$$H_1 : \sigma_1^2 > \sigma_2^2$$

Step 3: Defining the test statistic

Because the study is related to the differences of two population variances, the test statistic to be evaluated is F_c. It is defined as $F_c^2 = \frac{s_2^2}{s_1^2}$.

Step 4: Defining the significance level

Because the level of significance α is given as 0.01, according to F − table, $F_t(\alpha, v_1, v_2) = F_t(0.01, 10, 9) = 5.257$. We see that if F_c of the sample statistic satisfies the condition, $F_c \leq 5.257$, then we are 95% confident that H_0 is true ($P(F_c \leq 5.257) = 0.99$).

Step 5: Evaluate F_c

$$F_c = \frac{s_2^2}{s_1^2} = \frac{264}{250} = 1.056$$

Step 6: Statistical decisions

Since $F_c = 1.056$ lies in the acceptance region ($1.056 \leq 5.257$), according to the decision rule, we accept H_0.

Step 7: Conclusion

We conclude that at 5% level of significance, the differences between the two population variances are not significant.

Example:

Two samples are drawn from two normal populations. From the following data, test whether the two samples have the same variance at 5% level.

Sample 1:	60	65	71	74	76	82	85	87		
Sample 2:	61	66	67	85	78	63	85	86	88	91

Step 1: Given the data, $n_1 = 8$; $n_2 = 10$

Sample 1, *X*		***X*²**	**Sample 2, *Y***	***Y*²**
60		3600	61	3721
65		4225	66	4356
71		5041	67	4489
74		5746	85	7225
76		5776	78	6084
82		6724	63	3969
85		7225	85	7225
87		7569	86	7396
			88	7744
			91	8281
Total	600	45636	770	60490

$$s_1^2 = \frac{\sum_{i=1}^{n_1} x_i^2}{n_1} - \left(\frac{\sum_{i=1}^{n_1} x_i}{n_1}\right)^2 = \frac{45636}{8} - \left(\frac{600}{8}\right)^2 = 5704.5 - 5625 = 79.5$$

$$S_1^2 = \left(\frac{n_1}{n_1 - 1}\right) S_1^2 = \left(\frac{8}{7}\right)(79.5) = 90.86$$

Similarly proceeding, we have

$$s_2^2 = \left(\frac{60490}{10}\right) - \left(\frac{770}{10}\right)^2 = 120$$

$$S_2^2 = \left(\frac{n_2}{n_2 - 1}\right) s_2^2 = \left(\frac{10}{9}\right)(120) = 133.33$$

$$n_1 = 8;\ S_1^2 = 90.86,\ \ n_2 = 10,\ S_2^2 = 133.33$$

$$v_1 = n_1 - 1 = 8 - 1 = 7;\ v_2 = n_2 - 1 = 10 - 1 = 9$$

Let $\sigma_1{}^2$ and $\sigma_2{}^2$ be the two population variances.

Step 2: Framing the hypothesis

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq > \sigma_2^2$$

Step 3: Defining the test statistic

Because the study is related to the difference of two population variances, the test statistic to be evaluated is F_c. It is defined as, $F_c = \frac{S_2^2}{S_1^2}$.

Step 4: Defining the significance level

The level of significance α is given as 5%. Because it is a two- sided test, according to the F-table, $F_t(\alpha/2, v_2, v_1) = F_t(0.025,9,7) = 4.8232$.

We see that, if F_c of the sample statistic satisfies the condition, $F_c \leq 4.8232$, we are 95% confident that H_0 is true ($P(F_c \leq 4.8232) = 0.95$).

Step 5: Evaluate F_c

$$F_c = \frac{S_2^2}{S_1^2} = \frac{133.33}{90.86} = 1.467$$

Step 6: Statistical decisions

Since $F_c = 1.11$ lies in the acceptance region ($1.467 \leq 4.8232$), according to the decision rule, we accept H_0.

Step 7: Conclusion

We conclude that at the 5% level of significance with (10, 9) degrees of freedom, the difference between the two population means are not significant.

13.20 Analysis of Variance (ANOVA)

ANOVA is a technique whereby the total variation present in a set of data is partitioned into many components. Associated with each of these components is a specific source of variation, so that in the analysis, it is possible to ascertain the magnitude of the contribution of each of these sources to the total variation.

This technique was introduced and further developed by R. A. Fisher from 1912 to 1962. It had a tremendous influence on modern statistical thought. Fisher defined ANOVA as 'the separation of the variance ascribable to one group of causes from the variance ascribable to other groups'.

It is most often used to analyse data derived from designing experiments.

We use ANOVA to estimate the test hypothesis about both the population means and variances. In this chapter, we are going to deal only with testing hypothesis about the population means and the conclusions that depend on the magnitudes of the observed variances.

The valid use of ANOVA depends on a set of fundamental assumptions.

- Samples are selected randomly from the populations.
- All the populations were in which the samples are randomly selected follow normal distribution.
- The variances of all the populations are equal.

It is classified into two types:

- One-way classification
- Two-way classification.

One-way classification
Observations are based on 1 criterion (factor).

> **Example:**
>
> Consider the yield on 12 plots of land in three samples, each containing 4 plots. Each sample uses different varieties of fertilizers, namely, Brands F_1, F_2, and F_3, respectively.

Two-way classification
In this type of classification, the statistical data are classified according to two different criteria.

> **Example:**
>
> Consider the yield on 12 plots of land in 3 samples, each containing 4 plots. Each sample uses different kinds of seeds, S_1, S_2, and S_3 and different types of fertilizers F_1, F_2, and F_3, respectively.

Yield from Sample Different Plants Fertilizer Seeds

One-way analysis	A	F_1	S_1
	B	F_2	S_2
	C	F_3	S_3

Yield from Sample Different Plants Fertilizer Seeds

Two-way analysis	A	F_1	S_1
	B	F_2	S_2
	C	F_3	S_3

Notations

SST : Total sum of squares of deviation

SSB : Sum of squares of deviation between the samples

SSF : Sum of squares of devotion within the samples

CF : Correction factor

Consider m number of samples.

Sample 1	**Sample 2**	**…**	**Sample *m***
X_1	X_2	…	X_m
X_{11}	X_{21}		X_{m1}
X_{12}	X_{22}		X_{m2}
…	…		…
X_{1n_1}	X_{2n_2}		X_{mn_m}
Size: n_1	Size: n_2	…	Size: n_m

NOTE: The values of n_1, n_2, … , n_m need not be equal.

$$N = n_1 + n_2 + \ldots + n_m$$

Find the sum and square of sum values of the individual samples.

X_1	X_1^2	X_2	X_2^2	…	X_m	X_m^2
X_{11}	X_{11}^2	X_{21}	X_{21}^2	…	X_{m1}	X_{m1}^2
X_{12}	X_{12}^2	X_{22}	X_{22}^2	…	X_{m2}	X_{m2}^2
…	…	…	…	…	…	…
$X_{1[n1]}$	$X^2_{1[n1]}$	$X_{2[n2]}$	$X_{2[n2]}^2$	…	$X_{m[nm]}$	$X^2_{m[nm]}$
S_1 = Sum	SS_1 = Sum	S_2 =Sum	SS_2 = Sum	…	S_m = Sum	SS_m =Sum

$$T = \text{Total} = S_1 + S_2 + \ldots + S_m$$

$$CF = \frac{T^2}{N}$$

$$SST = (SS_1 + SS_2 + \ldots + SS_m) - CF.$$

$$SSB = \left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} + \ldots . + \frac{Sm^2}{n_m}\right) - CF$$

$$SSW = SST - SSB$$

Enter all the values into the ANOVA table.

Source of Variation	**Sum of Squares (SS)**	***df***	**Means Squares (MS)**	F_c
Between	*SSB*	$v_1 = m{-}1$	$MSB = \frac{SSB}{m-1}$	$F_c = \frac{SSB}{SSW}; if \;\; SSB > SSW(or)$
Within	*SSW*	$v_2 = N{-}m$	$MSW = \frac{SSW}{N-m}$	$F_c = \frac{SSW}{SSB}; if \;\; SSW > SSB$

Based on F_c and $F_t(\alpha,v_1, v_2)$ or $F_t(\alpha,v_2,v_1)$ conclude.

13.20.1 One-Way Classification

Example:

The following data give the yields on 12 plots of land in 3 samples, each of 4 plots, under 3 varieties of fertilizers A, B, and C:

A	**B**	**C**
25	20	24
22	17	26
24	16	30
21	19	20

Is there any significant difference in the average yields of land under the three varieties of fertilizers? Assume the significance level as 0.05.

Step 1: Consider the data given:

A	**B**	**C**
25	20	24
22	17	26
24	16	30
21	19	20

NOTE: Whenever the given values are larger, then we can subtract a common value (A) from each entry and the value of A can be selected arbitrarily.

Let us subtract the value 15 from all the entries. Hence, the modified data can be given as

Sample A	Sample B	Sample C
10	5	9
7	2	11
9	1	15
6	4	5

Here, $m = 3$; $n_1 = 4$; $n_2 = 4$; $n_3 = 4$.

$$N = n_1 + n_2 + n_3 = 12.$$

Let μ_1, μ_2, and μ_3 be the means of the three populations.

Step 2: Framing the hypothesis

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: at least 1 equality does not hold

Step 3: Defining the test statistic

Because the number of samples is more than 2, and the study is related to equality of population means, we make use of ANOVA with one-way classification. The test statistic to be evaluated is F_c. It is defined as,

$$F_c = \frac{MSB}{MSW}; \text{ if } MSB > MSW \text{ [or]}$$

$$F_c = \frac{MSB}{MSW}; \text{ if } MSW > MSB.$$

Step 4: Defining the significance level

Because the level of significance α is not given, let us assume that $\alpha = 0.05$. According to the F-table,

$$F_t(\alpha, v_1, v_2) = F_t(0.05,2,9) = 4.26$$

$$F_t(\alpha, v_2, v_1) = F_t(0.05,9,2) = 19.4$$

See that if F_c of the sample statistic satisfies the condition.

$F_c \leq F_t$, then we are 95% confident that H_0 is true ($P(F_c \leq F_t) = 0.95$).

Step 5: Evaluate SSB, SSW, and F_c

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
10	100	5	25	9	81
7	49	2	4	11	121
9	81	1	1	15	225
6	36	4	16	5	25
Total =32	266	12	46	40	452

$$T=\sum_{i=1}^{4}\times 1j+\sum_{i=1}^{4}\times 2j+\sum_{i=1}^{4}\times 3j = 32+12+40+=84;\ T=84 \text{ and } N=12$$

$$\text{Correction factor } (CF)=\frac{T^2}{N}=\frac{(84)^2}{12}=558;\ CF=558$$

$$\text{Total sum of square deviation } (SST)=\left(\sum_{j=1}^{4}\times 1j^2+\sum_{j=1}^{4}\times 2j^2+\sum_{j=1}^{4}\times 3j^2\right)-CF$$

$$SST = 266 + 46 + 452 - 588$$

$$SST = 176$$

Sum of the square deviation between the samples (SSB)

$$=\left(\frac{\left(\sum_{j=1}^{4}\times 1j\right)^2}{n_1}+\frac{\left(\sum_{j=1}^{4}\times 2j\right)^2}{n_2}+\frac{\left(\sum_{j=1}^{4}\times 3j\right)^2}{n_3}\right)-CF$$

$$=\left(\frac{(32)^2}{4}+\frac{(12)^2}{4}+\frac{(40)^2}{4}\right)-588$$

$$= (256 + 36 + 400) - 588$$

$$SSB = 104$$

Sum of the squares with in the sample (SSW) = $SST - SSB$

$$SSW = 176 - 104 = 72$$

Enter all the value in the ANOVA table

Source of Variation	**Sum of Squares (SS)**	***df***	**Mean Square**	***F***
Between	$SSB = 104$	$v_1 = m - 1 = 3 - 1 = 2$	$\frac{SSB}{m-1}=\frac{104}{2}=52$ $MSB=52$	$F=\frac{MSB}{MSW}$
Within	$SSW = 72$	$v_2 = N - m = 12 - 3 = 9$	$\frac{SSW}{N-m}=\frac{72}{9}=8$ $MSW=8$	$\{MSB > MSW\}$ $F:\frac{52}{8}=6.5$

$$\because MSB > MSW, \text{ use } F_t = (\alpha, v_1, v_2) = F_t(0.05,2,9) = 4.26$$

$$F_c = 6.5$$

Step 6: Statistical decisions

Since $F_c = 6.5$ lies in the rejection area (6.5 > 4.26), according to the decision rule, we reject H_0.

Step 7: Conclusion

We conclude that at the 5% level of significance with (2, 9) *df*, there is a significant difference between the three population means.

Example:
The Amirt Merchandising Company wishes to test whether its three salesmen A, B, and C tend to make sales of the same size or whether they differ in their sales. During the last week, there have been 14 sales calls: A made 5 calls; B made 4 calls; and C made 5 calls. Following are the weekly sales record of the three salesmen:

A	B	C
($)	($)	($)
300	600	700
400	300	300
300	300	400
500	400	600
0	—	500

Perform the analysis and draw your conclusions.

Step 1: Consider the given set of data,

Here, $m = 3$, $n_1 = 5$, $n_2 = 4$, $n_3 = 5$.

$$N = 5 + 4 + 5 = 14$$

$$N - m = 14 - 3 = 11$$

$$m - 1 = 3 - 1 = 2$$

Let us divide all the items by 100 to reduce the values given.

NOTE: This process is purely optional.

Hence, the modified value is

X_1	X_2	X_3
3	6	7
4	3	3
3	3	4
5	4	6
0	—	5

Let μ_1, μ_2, and μ_3 be the means of the three populations.

Step 2: Framing the hypothesis

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: at least 1 equality does not hold

Step 3: Defining the test statistic

Because the number of samples is more than 2, and the study is related to equality of population means, we make use of ANOVA with one-way classification. The test statistic to be evaluated is F_c. It is defined as,

$$F_c = \frac{MSB}{MSW};\ \text{if } MSB > MSW$$

or

$$F_c = \frac{MSB}{MSW};\ \text{if } MSW > MSB.$$

Step 4: Defining the significance level

Because the level of significance α is not given, let us assume that $\alpha = 0.05$. According to the F-table,

$$F_t(\alpha,v_1,v_2) = F_t(0.05,2,11) = 3.9823$$

$$F_t(\alpha,v_2,v_1) = F_t(0.05,11,2) = 19.4055$$

See that if F_c of the sample statistic satisfies the condition, $F_c \le F_t$, then we are 95% confident that H_0 is true ($P(F_c \le F_t) = 0.95$).

Step 5: Evaluate *SSB*, *SSW*, and F_c.

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
3	9	6	36	7	49
4	16	3	9	3	9
3	9	3	9	4	16
5	25	4	16	6	36
0	0	—	—	5	25
Total 15	59	16	70	25	135

$$T = \sum_{i=1}^{5} \times 1j + \sum_{i=1}^{4} \times 2j + \sum_{i=1}^{5} \times 3j = 15 + 16 + 25 = 56\ T = 56 \text{ and N} = 14$$

$$CF = \frac{T^2}{N} = \frac{(56)^2}{14} = 224$$

$$\text{Total sum of square deviation } (SST) = \left(\sum_{j=1}^{5} \times 1j^2 + \sum_{j=1}^{4} \times 2j^2 + \sum_{j=1}^{5} \times 3j^2 \right) - CF$$

$$SST = (57 + 70 + 135) - (224) = 40$$

$$SST = 40$$

Sum of the square deviation between the samples (*SSB*)

$$= \left(\frac{\left(\sum_{j=1}^{5} \times 1j \right)^2}{n_1} + \frac{\left(\sum_{j=1}^{4} \times 2j \right)^2}{n_2} + \frac{\left(\sum_{j=1}^{5} \times 3j \right)^2}{n_3} \right) - CF$$

$$= \left(\frac{15^2}{5} + \frac{16^2}{4} + \frac{25^2}{5} \right) - 224$$

$$SSB = (45 + 64 + 125) - 224$$

$$SSB = 10$$

$$SSW = \text{SST} - \text{SSB} = 40 - 10 = 30$$

Enter all the values in the ANOVA table:

Source of Variation	**SS**	***Df***	**MS**	**F**
Between	$SSB = 10$	2	$MSB = 10/2 = 5$	$F_c = \frac{MSB}{MSW}$
Within	$SSW = 30$	11	$MSW = 30/11 = 2.73$	$F_c = \frac{5}{2.73} = 1.83$

Since $MSB > MSW$, use $F_t(\alpha, v_1, v_2) = F_t\ (0.05,2,11) = 3.9823$

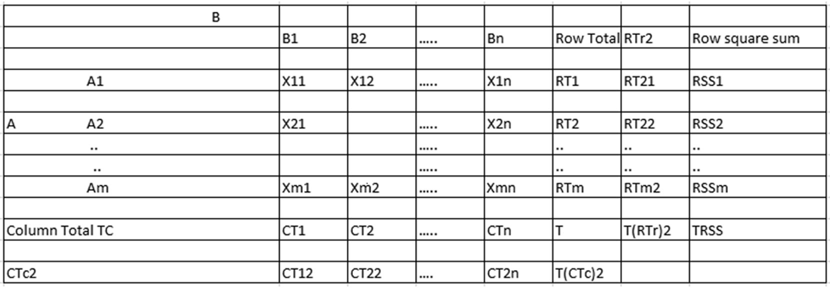
Step 6: Statistical decisions

Since $F_c = 1.83$ lies in the acceptance area ($1.83 \leq 3.9823$), according to the decision rule, we accept H_0.

Step 7: Conclusion

We conclude that at the 5% level of significance with (2,11) *df*, there is no significance difference between the three population means.

13.20.2 Two-Way Classification



B							
	B1	B2	.....	Bn	Row Total	RTr2	Row square sum
A1	X11	X12	.....	X1n	RT1	RT21	RSS1
A A2	X21		.....	X2n	RT2	RT22	RSS2
..			.....		..	..	..
..			.....		..	..	..
Am	Xm1	Xṁ2	.....	Xmn	RTm	RTm2	RSSm
Column Total TC	CT1	CT2	.....	CTn	T	T(RTr)2	TRSS
CTc2	CT12	CT22	....	CT2n	T(CTc)2		

$$RSS_i = j\text{th Row square sum;} = \sum_{j=1}^{n} x_{ij}{}^2 \; i = 1,4,\ldots..m.$$

$$RT_i = \text{Sum of all the elements in the } i\text{th row.}$$

$$CT_j = \text{Sum of all the elements in the } j\text{th column.}$$

$$T(RT_i)^2 = \sum_{i=1}^{m} RT_i^2$$

$$T(CT_j)^2 = \sum_{i=1}^{n} CT_j^2$$

$$T = \sum_{i=1}^{m} RT_i \; = \sum_{j=1}^{n} CT_j$$

Alternative method

Notations :
CF: Correction Factor
SSC : The sum of squares between the columns/variance between the column.
SSR : The sum of squares within the columns/variance between the rows.
TRSS : Sum of squares of individual element.

$$CF = T^2/N;\; SSC = \{T(RT_i)^2/r\} - CF$$

$$df = c\text{–}1;\; SSR = \{T(CT_j)^2/c\} - CF$$

$$df = r\text{–}1;\; SST = TRSS - CF$$

$$df = N - 1;\; SSE = SST - (SSC + SSR)$$

$$df = (r - 1)(c - 1)$$

Particulars	***df***
SST	N–1
SSC	c–1
SSR	r–1
SSE	(r–1)(r–1)

$$CF = \frac{T^2}{N};\text{ where } T \text{ refers the grand total.}$$

$$T = \sum_{i=1}^{m} r_i = \sum_{j=1}^{n} c_j$$

$$SSC = \frac{\left(\sum_{j=1}^{n} \times 1j\right)^2}{r} + \frac{\left(\sum_{j=2}^{n} \times 2j\right)^2}{r} + \ldots.. + \frac{\left(\sum_{j=1}^{n} \times mj\right)^2}{r} - CF$$

$$df = c–1 = n–1.$$

$$SSR = \frac{\left(\sum_{j=1}^{n} \times 1j\right)^2}{C} + \frac{\left(\sum_{j=1}^{n} \times 2j\right)^2}{C} + \ldots.. + \frac{\left(\sum_{j=1}^{n} x_{mj}\right)^2}{C} - CF$$

$$df = r–1$$

$$SST = \sum_{n-1}^{m} \sum_{j=1}^{n_i} (x_{ij})^2 - CF$$

$$df = N–1$$

$$SSE = SST–(SSR + SSC)$$

$$df = (N - 1)–(c - 1) + (r–1))$$

$$df = N - 1 - \mathrm{c} + 1 - 0 + 1 = (N + 1)–(c + r)$$

Enter the values in to the ANOVA table (Two-Way)

Source of Creation	Sum of Squares (SS)	*df*	Mean Square (MS)	F Statistic	Nr. *df* V_1	Dr. *df* V_2	$F_t(\alpha)$	Decision
Between Columns	*SSC*	*c* – 1	$MSC = \frac{SSC}{c-1}$	$F_C = \frac{MSC}{MSE}$ *or*	c–1	*k*	$F_c \le F_t$	Accept
				$F_C = \frac{MSE}{MSC}$	*k*	*c*–1	$F_c > F_t$	Reject
Between Rows	*SSR*	*r*–1	$MSR = \frac{SSR}{r-1}$	$F_C = \frac{MSR}{MSE}$ or	*r*–1	*k*	$F_c \le F_t$	Accept
				$F_C = \frac{MSE}{MSR}$	*k*	*r*–1	$F_c > F_t$	Reject
Residual	*SSE*	*k*	$MSE = \frac{SSE^2}{k}$					

NOTE 1: $k = (c–1)(r–1)$

NOTE 2: $F_c = \frac{MSC}{MSE}$; if $MSC > MSE$;

$F_c = \frac{MSE}{MSC}$; If $MSE > MSC$.

Similarly, one can find the second set of F_c value. We must conclude based on the decision.

Example:

A company appoints four salesmen S_1, S_2, S_3, and S_4 and observes their sales in three seasons: Summer, Winter, and Monsoon. The data regarding the sales figures unified as 1unit = 1 lakh of rupees.

Season/ Salesmen	S_1	S_2	S_3	S_4	**Season Total (Rs).**
Summer	36	36	21	35	128
Winter	28	29	31	32	120
Monsoon	26	28	29	29	112
Salesmen Total	90	93	81	96	360

Verify using two-way ANOVA if there is any significant difference in total sales among the 4 salesmen and also if there is any significant difference in sales with respect to three seasons.

Step 1: Consider the data given, let us subtract 20 from all the entries. The modified data (coded data)

Season/ Salesmen	S_1	S_2	S_3	S_4	**Total**
Summer	16	16	1	15	48
Winter	8	9	11	12	40
Monsoon	6	8	9	9	32
Salesman Total	30	33	21	36	120

$$m = 3; c = 4; r = 3; N = 12; T = 120.$$

Step 2: Framing the hypothesis

H$_0$: There is no significant difference between the mean sales based on different salesmen or different seasons.

H$_1$: There is a significant difference between the mean sales based on different salesmen or different seasons.

Step 3: Defining the test statistic

Because the number of samples is more than 2, we have to make use of ANOVA. Also, since the study is based on both the parameters, it is a two-way classification.

$$(FC)_c = \frac{MSC}{MSW}[\text{OR}]\frac{MSE}{MSC}$$

$$(FC)_r = \frac{MSR}{MSE}[\text{OR}]\frac{MSE}{MSR}$$

Step 4: Defining the significance level

Because the level of significance is not given, let us assume that $\alpha = 0.05$.

According to the F-table,

$$(F_t)_c\ (\alpha,v_1,v_2) = (F_t)_c(0.05,6,3) = 8.94$$

$$(F_t)_r\ (\alpha,v_1,v_2) = (F_t)_r(0.05,6,2) = 19.3$$

We see that if F_c of the sample statistic satisfies the condition $(Fc)_c \leq .94$ and $(Fc)_r \leq 19.3$ *in both the cases we are 95% confident that* H_0 *is true.*

Step 5: Evaluate *CF*, *SSC*, *SSR*, *SST*, and *SSE*.

$$CF = \frac{T^2}{N} = \frac{120^2}{12} = 1200$$

$$SSR = \frac{48^2}{4} + \frac{40^2}{4} + \frac{32^2}{4} - 1200 = 1232 - 1200 = 32$$

$$SSR = 32;\ df = r-1 = 2$$

$$SSC = \frac{30^2}{3} + \frac{33^2}{3} + \frac{21^2}{3} + \frac{36^2}{3} - 1200$$
$$= 1242 - 1200 = 42$$

$$SSC = 42;\ df = c-1 = 3$$

$$SST = 16^2 + 8^2 + 6^2 + 16^2 + 9^2 + 8^2 + 1^2 + 11^2 + 9^2 + 15^2 + 12^2 + 9^2 - 1200$$

$$SST = 1410 - 1200 = 210$$

$$SST = 210;\ df = N-1 = 11$$

$$SSE = SST-(SSR + SSC) = 210-42-32 = 136$$

$$SSE = 136;\ df = (N-1) - \{(C-1) + (r-1)\} = 6$$

Enter the values into the two-way classification table.

Source of Variation	**Sum of Squares (*SS*)**	***df***	**Mean Square**	F_c	***Nr df***	***Dr df***	**Table Value**	**Decision**
Between columns	$SSC = 42$	3	$\frac{42}{3} = 14$ $MSC = 14$	$\frac{MSE}{MSC} = \frac{22.67}{14}$ $F_c = 1.619$	6	3	$F_t = 8.94$	1.619 < 8.94 Accept H_0
Between Rows	$SSR = 32$	2	$\frac{32}{2} = 16$ $MSR = 16$	$\frac{MSE}{MSR} = \frac{22.67}{16}$ $F_c = 1.417$	6	2	$F_t = 19.3$	1.417 < 19.3 Accept H_0
Residual	$SSE = 136$	6	$\frac{136}{6} = 22.67$ $MSE = 22.67$					

Step 6: Statistical Decisions

1. Since $(F_c)_c = 1.619 < 8.94$, according to the decision rule, we accept H_0.
2. Since $(F_c)_r = 1.417 < 19.3$, according to the decision rule, we accept H_0.

Step 7: Conclusion

1. We conclude that at the 5% level of significance with $df(6,3)$, there is no significant difference between the salesmen.
2. We conclude that at 5% level of significance with $df(6,2)$, there is no significant difference between the seasons.

Example:

Consider the final examination scores secured by the students of different disciplines learned on 3 different instructional methods.

		Different Way of Teaching			
		Lecturer M_1	**Cases M_2**	**Problems and Discussion M_3**	**Total**
Different Discipline	Eng, D_1	61	60	77	218
	Business D_2	59	79	76	214
	Economics D_3	56	78	68	202
	Mathematics D_4	54	66	63	183
	Statistics D_5	45	72	66	183
	Total	275	375	350	1000

Test the null hypothesis that there is no difference in final examination scores among the 3 methods of instruction and 5 different disciplines. Test at 5% level of significance.

Step 1: Consider the data given.

Let us subtract 45 from all the entries the modified data (coded data) is

	M_1	**M_2**	**M_3**	**T_r**	**RSS**	**Tr^2**
D_1	16	35	32	83	2505	6889
D_2	14	34	31	79	2313	6241
D_3	11	33	23	67	1739	4489
D_4	9	21	18	48	846	2304
D5	0	27	21	48	1170	2304
Tc	50	150	125	325		
Tc^2	2500	22500	15625	40625	8573	22227

Here

$$T = 317$$

$$C = 3;\ df = c–1 = 2$$

$$r = 5;\ df = r - 1 = 4$$

$$N = c \times r = 15;\ df = N - 1 = 14.$$

Step 2: Framing the hypothesis

H$_0$: There is no significance difference between the final examination score based on different methods of teaching and different disciplines of students.

H$_1$: There is a significant difference between the final examination score based on different methods of teaching and different disciplines of students.

Step 3: Defining the test statistic

Because the number of samples is more than 2, we make use of ANOVA. Also, since the study is based on both the parameters, different discipline of students and different teaching methods, it is a two-way classification.

1. $(F_c)_c = \dfrac{MSC}{MSE}[\text{or}]\dfrac{MSE}{MSC}$
2. $(F_c)_r = \dfrac{MSR}{MSE}[\text{or}]\dfrac{MSE}{MSR}$

Step 4: Defining the significance level

Because the level of significance α is given as 0.05, according to the table,

$$(F_t)_c(\alpha,v_1,v_2) = (F_t)_c(0.05,2,8) = 4.46$$

$$(F_t)_r(\alpha,v_1,v_2) = (F_t)_r(0.05,4,8) = 3.84$$

We see that if F_c of the sample statistic satisfies the condition $(F_c)_c \leq 4.46$ and $(F_c)_r \leq 3.84$

In both the cases, we are 95% confident that H$_0$ is true.

Step 5: Evaluate *CF*, *SSC*, *SSR*, *SST* and *SSE*.

$$CF = \frac{T^2}{N} = \frac{325^2}{15} = 7041.67$$

$$SSR = \frac{\sum Tr^2}{c} - CF$$

$$SSR = \frac{22227}{3} - CF = 7409 - 7041.67 = 367.33$$

$$df = 5–1 = 4$$

$$SSC = \frac{\sum Tc^2}{5} - CF = \frac{40625}{5} - 7041.67 = 8125 - 7041.67 = 1083.33$$

$$SSC = 1083.33$$

$$df = 3–1 = 2.$$

$$SST = \Sigma RSS - CF = 8573\text{–}7041.67 = 1531.33$$

$$df = 15\text{–}1 = 14$$

$$SSE = SST - (SSC + SSR) = 1531.33 - (1083.33 + 367.33) = 8067$$

$$df = 14 - (2 + 4) = 8$$

Enter the values in to the two-way analysis table.

Source of Variation	Sum of Squares (*SS*)	*df*	Mean Square	F_c	N_r *df* V_1	*Dr df* V_2	*Ft* $\alpha = 0.05$	Decision
Between columns	$SSC = 1083.33$	2	$MSC = 541.67$	$\frac{MSC}{MSE}$ $= \frac{541.67}{10.08}$ $= 53.74$	2	8	4.46	53.74 > 4.46 Reject H_0
Between Rows	$SSR = 367.33$	4	$MSR = 91.83$	$\frac{MSR}{MSE}$ $= \frac{91.83}{10.08}$ $= 53.74$	4	8	9.11	9.11 > 3.84 Reject H_0
Residual	$SSE = 80.67$	8	$MSE = 10.08$					

Step 6: Statistical decisions

1. Since $(Fc)_c = 53.74 > 4.46$, according to the decision rule, we reject H_0.
2. Since $(Fc)_r = 9.11 > 3.84$, according to the decision rule, we reject H_0.

Step 7: Conclusion

1. We conclude that at the 5% level of significance with (2,8) *df*, there is a significant difference in the final examination scores for the different instructional methods.
2. We conclude that at the 5% level of significance with (4,8) *df*, there is a significant difference in the final examination scores for the students with different disciplines.

Exercise 13

1. A filling machine at a soft drink factory is designed to fill bottles of 200 mL with a SD of 10 mL. A random sample of 50 filled bottles was taken and the average volume of soft drink was computed to be 198 mL per bottle. Test the hypothesis that the mean volume of soft drink per bottle is not less than 200 mL at 5% level of significance.
2. The sales manager of a large company conducted a sample survey in states A and states B taking 400 sample salesmen in each case. The results were:

	State A [$]	State B [$]
Average sales	2500	2200
SD	400	550

Test whether the average sales is the same in the two states at 1% level of significance.

3. A survey of television audience in a big city revealed that 50 out of 200 males and 80 out of 250 females liked a particular nightly program. Test the hypothesis at 5% level of significance whether there is a real difference of opinion about the program between male and female audiences.
4. The mean yield of two sets of plots, and their variability is as given below. Examine:
   a. Whether the difference in the mean yields of the two sets of plots is significant and
   b. Whether the difference in the variability in yields is significant.

	Set of 40 Plots	Set of 60 Plots
Mean yield per plot	1258 kg	1243 kg
SD per plot	34	28

5. The mean life of a sample of 100 electric bulbs produced by a company is found to be 1570 hours with a SD of 120 hours. If μ is the mean lifetime of all the bulbs produced by the company, test the hypothesis $\mu = 1600$ hours against the alternative hypothesis $\mu \neq 1600$ hours, using a level of significance of 0.05.
6. In a sample of 400 parts manufactured by a factory, 30 items are found to be defective. The company, however, claimed that only 5% of their product is defective. Verify whether the claim is true.
7. In Trichy district, 450 persons were considered regular consumers of tea out of a sample of 1000 persons. In another district, Pudukkottai, 400 were regular consumers of tea out of a sample of 800 persons. Do these facts reveal a significant difference between the two districts as far as a tea-drinking habit is concerned? Test this at 1% level of significance.
8. Consider two different groups of people each containing with 100 members affected by a specific disease. An injunction is given to group 1 but not to group 2. It is found that in group 1 and group 2, 75 and 65 members, respectively, recover from the disease. Test the hypothesis with 5% level of significance that the injunction helps to cure the disease.
9. Of shoppers at Market A, 400 are randomly chosen. Their mean weekly expenditure on purchasing is found to be $250 with a SD of $40. These figures are $220 and $55, respectively in Market B with the size of 400 shoppers selected at random. Test at 5% level of significance whether the average weekly expenditures of the two populations of shoppers are equal.
10. A company claims that its batteries are superior to those of its competitor on the basis of a study that showed that a sample of 50 of its batteries had an average lifetime of 75 hours of continuous use with a sample of SD of 7 hours, whereas a sample of 30 of the competitor's batteries had an average lifetime of 70 hours of continuous use with

a sample SD of 5 hours. Test at the 5% significance level the null hypothesis $\mu_1 = \mu_2$ against the alternative hypothesis $\mu_1 > \mu_2$ to verify whether this claim is justified.

Are the mean hourly wage rates in manufacturing industries the same in the two cities? Use the level of significance as 1%.

11. A fast-food trade association has published a statement that Bhavanasree Corporation has a market share of no more than 30% of the fast-foods business. Bhavanasree's Management, however, believes that its market share is greater than 30%. Consequently, the company commissioned a survey of 400 customers to determine the proportion who purchased fast foods from Bhavanasree's. This proportion turned out to be 140/400. Based on the survey, who is right? Test at the 5% level of significance.
12. In a city, a sample of 1000 people were taken and out of them 540 are vegetarians and the rest are not. Can we say that both habits of eating (vegetarian or non-vegetarian) are equally popular in the city at 1% level of significance?
13. A company claims that its batteries are superior to those of its competitor on the basis of a study that showed that a sample of 50 of its batteries had an average lifetime of 75 hours of continuous use with a sample standard deviation of 7 hours, while a sample of 30 of the competitor's batteries had an average lifetime of 70 hours of continuous use with a sample standard deviation of 5 hours. Test, at 5% significance level the null hypothesis H_0: $\mu_1 - \mu_2 = 0$; against the alternative hypothesis H_1: $\mu_1 - \mu_2 \neq 0$ to verify whether this claim is justified.
14. A manufacturing firm claims that its brand A product outsells its brand B product by 8%. If it is found that 42 out of a sample of 200 persons prefer brand A and 18 out of another sample of 100 persons prefer brand B. Test whether the 8% difference is a valid claim. (Refer to Example 15; use the formula given in the special note.)
15. The diameter of a particular bearing supplied by Vendor 1 and Vendor 2 follows a normal distribution. The quality manager of the firm buying the bearings believes that the diameter of the bearings supplied by vendor 1 is not different from that of the bearings supplied by vendor 2. So, the quality assistant under him selected 16 bearings from the supply of the vendor 1 and found that mean and variance of the diameter as 38.5 mm and 2.5 mm. respectively. Similarly, he selected 20 bearings from the supply of the vendor 2 and found that mean and variance of the diameter as 40 mm and 4 mm, respectively. Verify the intuition of the manager at a significance level of 0.1.
16. Sandal powder is packed into packets by a machine. A random sample of 12 packets is drawn and their weights are found to be [in kilograms]: 0.49, 0.48, 0.49, 0.50, 0.51, 0.49, 0.48, 0.50, 0.51, and 0.48. Test if the average packing can be taken as 0.5 kg. [By considering the data given, first evaluate the values of mean and SD and then apply the formula for testing the hypothesis. Mean = 0.49; SD = 0.012.]
17. A production manager feels that the output rate of experienced employees is surely greater than that of new employees, but he does not expect the variability in output rates to differ for the two groups. In previous output studies, it has been shown that the average unit output per hour for the 20 employees with few years of experience at this work is 30 units per hour with variance of 28 units.

For a group of 20 new employees, the average output for the same type of work is 20 units per hour with a sample variance of 56 units. Does the variability in the output of the new group differ from that of the experienced group? Test the hypothesis at 0.05 level of significance [α].

18. Samples of two types of electric bulbs were tested for length of life and the following data were obtained:

Particulars	Type I	Type II
Number in the sample	8	7
Mean of the sample [hrs]	1134	1024
Standard deviation of the sample	40	35

Test whether the two types of bulbs have the same length of life.

19. Marketing personnel A and B are working for two different districts. A sample survey conducted by the company reveals the following results. State whether there is any significant difference in the average sales between the two salesmen:

	A	B
Number of sales	20	18
Mean sales [$]	170	205
SD [$]	20	25

20. A powder manufacturing company was distributing a particular brand of powder through a retail shops. The actual average sale per week per shop before the advertisement campaign was 140 dozen. After the advertisement campaign, a sample of 26 shops was taken and the average sale was found to be 147 dozen with SD of 16. Test whether the advertisement campaign is effective.

21. A certain diet newly introduced to each of the 12 cows results in the following increase in body weight:

6, 3, 8, –2, 3, 0, –1, 1, 6, 0, 5, and 4.

Test whether the diet is quite effective in increasing the weight of the cows.

22. To see whether silicon chip sales are independent of where US economy is in the business cycle, data have been calculated on the weekly sales of a firm and on whether the US economy was rising to a cycle peak, at a cycle peak, falling to a cycle peak or at a cycle trough. The results are

	Weekly	Chip	Sale	
	High	Medium	Low	Total
Eco. at peak	20	7	3	30
Eco. at trough	30	40	40	100
Eco. at rising	20	8	2	40
Eco. falling	30	5	5	40
Total	100	60	40	200

Since the null hypothesis at 0.10 significance level, what is your conclusion? [Given the tabulated value of the test statistic is 10.645.]

23. The following table gives the number of screws declared fit and unfit by three inspectors *X*, *Y*, and *Z*. Test the hypothesis that the proportion of screws declared unfit by the three inspectors are same.

Inspectors	***X***	***Y***	***Z***	**Total**
Fit screws	50	47	56	153
Unfit screws	5	14	8	27
total	55	61	64	80

24. Apply the chi-square test to find out whether the injection is quite effective with respect to the disease:

	Affected	**Not Affected**
Injection used	20	300
Injection not used	80	600

25. Out of 2000 members exposed to smallpox in a town, 450 were attacked. Among the people 365 were vaccinated and out of them 50 were affected. Test using chi-square that whether vaccination can be regarded as a good preventive medicine or not.

26. The following table gives the number of aircraft accidents that occurred during the various days of the week. Test whether the accidents are uniformly distributed over the week.

Days	**Mon**	**Tue**	**Wed**	**Thur**	**Fri**	**Sat**
No. of Accidents	14	18	12	11	15	14

[Note: expected no. of accidents on any day = Mean

Mean = total number of accidents/6 = 84/6 = 14]

27. 400 workers were selected at random from a district. Their mean income was $104.5 per month with SD of $25.20. Do you believe that the average income of the working community in the district is $150?

28. Memory capacity of 9 workers was tested before and after a course of meditation for a month. Based on the data given, state whether the course was:

Before Meditation	10	15	9	3	7	12	16	17	4
After Meditation	12	17	8	5	6	11	18	20	3

29. To compare the prices of a certain product in two cities, 10 retail shops were selected at random in each town. The price was noted as follows:

City I	61	63	56	63	56	63	59	56	44	61
City II	55	54	47	59	51	61	57	54	64	58

30. Based on the information given, find out whether the new treatment is comparatively superior to the conventional one.

	Favourable	Not Favourable	Total
Conventional	40	70	110
New	60	30	90
	100	100	200

31. In an investigation into the health and nutrition of two groups of children of different social status, the following results are obtained.

Social Status Health Status	Poor	Rich	Total
Below normal	130	20	150
Normal	102	108	210
Above normal	24	96	120
Total	256	224	480

Discuss the relation between the health and their social status.

32. Suppose that the Palmer Corporation is planning to purchase a component for its guidance system used in various commercial aircraft. The company wishes to have components that possess not only a long life but also a high degree of uniformity. From experience, the SD should not exceed 14 hours A sample of 25 components yields a mean life of 1500 hours of operation, which is considered to be adequate. But with a SD of 17 hours, does the population SD of component life exceed a SD of 17 hours?

    [Note: H_0: σ^2 <= 196; H_1: $\sigma^2 > 196$; χ^2 calculated value = $[n - 1]s^2/\sigma^2 = 35.38$; χ^2 table value with 5% level of significance and 24 $df = 36.415$]

33. The transit authority of India plans to purchase light bulbs for its subway system. The company wishes to have bulbs that possess not only a long life but also a high degree of uniformity. It decides on experience, that the variance should not exceed 200. A test of 20 bulbs of a certain make yields a mean life of 1000 hours, which is satisfactory, and a variance of 250. Do these bulbs satisfy the company's uniformity requirement at 5% level of significance?

    [Note: H_0: σ^2 <= 200; H_1: $\sigma^2 > 200$; χ^2calculated value = $[n–1]s^2/\sigma^2 = 23.75$; χ^2 table value with 5% level of significance and 19 $df = 30.14$]

34. A survey of buying habits was conducted last week by Market Surveys Inc. in Chennai and Trichy. In Trichy, 100 homemakers were interviewed, and it was found that they spend an average of $6000 on food per month with a SD of $1250, whereas in Chennai, 200 homemakers reported an average monthly expenditure of $6500 with a SD of $1750. Using a significance level of 5%, test the hypothesis that there is no difference in the average amount spent on food per month between homemakers in Trichy and Chennai.

35. Market Survey Inc. is contracted to determine if the variance of money spent per family for entertainment in city 1 is significantly greater than the variance of

money spent per family for entertainment in city 2. For the two populations, the amounts of money spent are independent and normally distributed random variables. The data are as follows:

City	Size	Variance
1	16	186
2	20	81

Test the hypothesis at 5% level of significance that there is no significant difference between the variances in the two cities.

36. Determine based on the sample data shown in the following table, whether the true proportion of voters favouring increased defence spending by the federal government is the same in all three areas of the country. Test at the 5% level of significance.

Areas	Number Favouring Increased Defence Spending	Number Not Favouring Increased Defence Spending	Total
East	310	190	500
Midwest	236	164	400
West	174	126	300
Total	720	480	1200

37. To demonstrate the safety of its cars, WV Motors wishes to determine whether there is an association between the number of accidents and make of car owned. Available records reveal the following frequency of accidents by car make:

Number of accidents

Car Make	0	1	2	>2
WV	38	40	18	4
Renault	45	36	14	5
Other	55	26	10	9

Is there any relationship between car make and number of accidents? Test at the 5% significance level.

38. The contingency table that follows summarizes the results obtained in a study conducted by Consumer Studies Inc. with respect to the performance of four competing brands of toothpaste

Number of Cavities	Brand A	Brand B	Brand C	Brand D	Total
0	9	13	17	11	50
1–5	63	70	85	82	300
>5	28	37	48	37	150
Total	100	120	150	130	500

Test the hypothesis that the variables "incidence of cavities" and "brand used" is independent at the 5% level.

39. The following table compares four different occupational groups with their opinions about business conditions in the coming year. The question is whether the four groups think alike or whether their opinions differ significantly at 5% level of significance.

    Opinion about the future of business conditions as expressed by four occupational groups

Occupation	**Much Better**	**Better**	**Same**	**Worse**	**Much Worse**	**Total**
Farmers	21	13	22	30	46	132
Bankers	11	17	12	17	12	69
Merchants	10	10	12	12	10	54
Doctors	28	26	25	28	31	138
Total	70	66	71	87	99	393

40. A random sample of 168 college professors was asked to express an opinion as to whether research, teaching, or total performance is the most important basis for academic promotion. The survey results are shown in the following table:

		Teaching Field		
	Sciences	**Professional**	**Arts**	**Total**
Research	32	17	17	66
Teaching	12	22	22	56
Total performance	12	22	12	46
Total	56	61	51	168

    Use chi-square test with a level of significance of 0.05 to test the hypothesis that the universe distribution of proportion of opinion is the same for all the faculty groups.

41. There are three main brands of a certain powder. A set of 12 sales is examined and found to be allocated among four groups [A, B, C, and D] and brands [I, II, and III] as shown below:

		Replications			
		Groups			
Brands		**A**	**B**	**C**	**D**
Factor I	I	32	35	31	30
Factor II	II	30	24	32	26
Factor III	III	26	27	25	30

    Check whether the factor Brand has a significant effect on the sales at α =0.05 using one-way ANOVA.

42. The R&D manager of an automobile company wishes to study the effect of 'Tire Brand' on the tread loss [in millimetre] of tires. Four tires from each of four different brands [A, B, C, and D] are fitted to four different cars using the completely randomized design. The data as per this design are presented below:

	Tire Brand		
A	**B**	**C**	**D**
12	14	12	14
15	17	19	21
18	12	20	25
10	19	23	20

Check whether the tire brand affects the tread loss of tires at a significance level of 5%.

43. The following table shows the lives in hours of four batches of electric bulbs.

Batches								
1	1600	1610	1650	1680	1700	1720	1800	
2	1580	1640	1640	1700	1750			
3	1460	1550	1600	1620	1640	1660	1740	1820
4	1510	1520	1530	1570	1600	1680		

Perform an analysis of variance and show that a significant test does not reject their homogeneity. [Table value $F_{0.05}$ [3, 22] = 3.05]

44. Four different types of training programs were used in training 12 athletes competing in the 400-m dash. Three athletes were assigned randomly to each training program for the purpose of comparing the effect of the training program on performance. Each athlete's performance times [in seconds] in the race is shown in the following table and were used as the measure of analysis. Test the null hypothesis that there were no differences among the mean times for the four types of training program, using the 5% significance level.

Training Program	**A**	**B**	**C**	**D**
Times	42	55	50	50
	46	56	40	63
	45	61	48	49

45. The sales [in thousands of rupees per month] of four brands of a product under 3 promotional strategies are given as follows. Test the null hypothesis that the average sales of the brands do no differ, using the 1% significance level.

Promotional Strategy	**Brand A**	**Brand B**	**Brand C**	**Brand D**
Newspaper advertising	77	69	73	82
TV advertising	73	63	75	79
Special discount	81	75	80	72

46. Use ANOVA for the following data and test whether the mean yields of the varieties are equal and test the equality of the block mean.

Varieties	Blocks			
	I	II	III	IV
A	4	8	6	8
B	5	5	7	8
C	6	7	9	5

47. The following data give the number of units produced per day by 4 workers A, B, C, and D using four machines M1, M2, M3, and M4.

	M1	M2	M3	M4
A	45	42	48	38
B	40	32	50	34
C	43	36	44	40
D	36	38	46	36

Appendix A: Answers to Exercise Problems

Exercise 4

Qn. No.	Mean	Median	Mode	SD	CV
1	57.15			12.15	21.25
2	8.35	8.6		0.85	10.13
3*	364.42; 0.46			20.59; 0.02	5.65%; 4.35%
4	21.25	18.32	8.08	15.55	73.2%
5*	291.5; 286			4.9; 18.12	1.68%; 6.34%
6	39.59	38.8	33.75	9.06	22.89%
7	21.25	18.32	7.08	15.55	73.18%
8*	5.12; 6.16			2.81;2.23	54.88%; 36.2%
9	1274.5		1239.5	719.53	56.43%
10*	210.45; 220.71			48.99; 66.58	23.28%; 30.17%
11	3.22	3.1	2.95	1.03	31.96%
12	338.32	347.56	345.5	89.31	26.4%
13	66.3	65.44	62.31	9.03	13.62%
14	18.77			12.77	68.02
15	1014.16			82.61	8.15%
16	10.74	10.82	11.17	5.06	47.1%
17	61.18	61.07	61	5.58	9.13%

Refer: "*"

- 3. The price of gold is having more variability.
- 5. The performance of type A ball is better than type B ball.
- 8. Model B has more uniformity, and it is the best.
- 10. Company X is more consistent than Company Y.

Exercise 5

1. Mean = 69.3; MD = 11.3; SD = 13.66; CV = 19.71%
2. CV-A = 4.83%; CV-B = 5.71; Firm B is having more variability in salary distribution; Firm B is paying more total salary than Firm A; when Firm A and Firm B are combined the average salary is Rs. 180.
3.

Mean	SD	CV	Remark
X: 88	11.33	12.88	
Y: 83.8	5.381	6.42	Sales person Y is more consistent
Z: 104.2	14.62	14.03	

4. Firm A: CV = 1.9; Firm B: CV = 2.32; Firm B has more variability in wage.

5. Price in city A is more stable than Price in city B. CV-A: 11.13; CV-B: 24.59
6. $Q_1 = 957.14$; $Q_3 = 1076.67$; QD = 59.765
7. MD = 10.4
8. SD = 2.77 lakhs

Exercise 6

1. Mean = 18.066; mode = 18.5; SD = 1.775; kurtosis = 2.55; skewness = –0.73
2. $Q_1 = 30$; $Q_3 = 70$
3. Mean = 18.066; mode = 18.39; SD = 1.775; kurtosis = 2.55; skewness = –0.55
4. Mean = 34.81; mode = 34.38; SD = 17.1; kurtosis = 2.16; skewness = 0.075
5. $\mu_4 = 45416.15$; $\mu_2 = 141.0673$; kurtosis = 2.28222
6. $\mu_4 = 302.9811$; $\mu_2 = 9.3966$; kurtosis = 3.431439
7. Mean = 147.2; SD = 7.208; kurtosis = 2.466987; skewness = 0.090

Exercise 7

Qn. No.	Mean x	Mean y	SDx	SDy	R	cov	b_{yx}	b_{xy}	*y* on *x*	*x* on *y*
1	42.5	687.5	11.46	78.70	–0.91	–818.75	–6.24	–0.13	$y = -6.244x +$ 952.62; when $x = 10$; $y = 1015$	$x = -0.13y +$ 133.38
2	50	566	28.28	235.93	0.95	6320	7.9	0.11	$y = 7.9x +$ 171	$x = 0.11y -$ 14.26
3					0.72					
4					$R_{12} = -0.21$	$R_{13} = 0.64$	$R_{23} = -0.3$		Judges 1 and 3 having nearest approach	
5	4	7			0.5					
6									$y = 1.12x - 5.8$; when $x = 70$; $y = 72.6$	$x = 0.57y + 26.71$
7	*x*-sales	*y* = adv exp.							$y = 0.135x + 0.6$ when $x = 60$ $y = 8.7$	$x = 6y + 4$ when $y = 10$ $x = 64$
8									$y = 0.284x - 2.64$ when $x = 50$ $y = 11.58$	$x = 3.24y + 12.86$ when $y = 20$ $x = 77.69$
9					–0.08					
10					0.10					
11									$y = 0.398x - 134.79$	

(*Continued*)

Qn. No.	Mean x	Mean y	SDx	SDy	R	cov	b_{yx}	b_{xy}	*y* on *x*	*x* on *y*
12	104.1	362.9	19.98	72.43	0.91	**1313.91**	3.29	0.5	$y = 3.29x + 20.18$	$x = 0.25y + 13.23$
13	225.08	182.42	84.84	68.43	0.96	**5545.88**	0.77	1.18	$y = 0.77x + 9$ when $x = 200$; $y = 163$	$x = 1.18y + 9.09$
14	23.78	26.63	7.12	7.7	0.87	48.07	0.93	0.81	$y = 0.93x + 4.52$	$x = 0.81\ y + 2.2$ when $y = 40$; $x = 34.6$
15	10.5	451.25	3.28	149.74	0.849	416.63	38.76	0.019	$y = 38.76x + 44.314$	$x = 0.019y + 2.116$
16	599.83	180.5	36.63	16.05	0.96	409.25	0.58	1.589	$y = 0.577x - 165.66$	$x = 1.59y + 313.05$
17	107.83	120.42	13.75	95.57	0.67	884.07	4.67	0.097	$y = 4.674x - 383.62$	$x = 0.096y + 96.18$

Exercise 8

1. (a) 1/6 (b) 2/9
2. 1/5
3. —
4. 0.9
5. 0.411
6. 0.5
7. 0.116, 0.326, 0.558
8. 0.6, 0.94
9. 0.88
10. 1/99, 14/33, 35/99
11. 0.45, 0.55, 0.05, 0.11
12. 1/114, 91/228, 137/228
13. 5/11, 53/66
14. 3/5
15. 97/120
16. 320
17. 0.5, 13/120
18. 0.184, 0.123, 0.0034, 0.659
19. 0.851
20. 0.175
21. 0.71
22. 0.5, 0.46
23. 0.01, 0.23

Exercise 9

1. E[x] = 15.75; he will purchase it.
2. 0.0002, 0.0006, 0.019, 0.085, 0.8952; E[x] = 3.725; P[loosing] = 0.8952; Actual loss = 6 – 3.725 = Rs. 2.275.
3. K = 0.11; 0.88; 0.22; 0.88
4. E[x] = 2; variance = 1

Exercise 10

1. 0.995, 0.79
2. 0.223, 0.1912
3. 0.062; 0.0892
4. 0.9802 [98020], 0.0196 [19604], 0.0002 [20]
5. 0.4331
6. 0.151, 0.089
7. 61,9

Exercise 11

1. (a) 0.1446; 15 approximately (b) 0.5066; 51 approximately (c) 0.0222; 3 approximately
2. (a) 0.0548 (b) 0.4772
3. (a) 0.7745 (b) 0.0228
4. (a) 0.2266 (b) 0.4181 (c) 0.1598
5. (a) 0.3446; 173 (b) 0.3811; 191
6. 0.93; 0.081; 0.47
7. 0.05262

Exercise 12

1. SE($\overline{x}$) = 0.1061; $Z_{0.05}$ = 1.96; $Z_{0.1}$ = 1.645

   Confidence interval with 5% level of significance: (19.792, 10.208);

   Confidence interval with 1% level of significance: (9.726, 10.274)

NOTE: Obviously when the level of significance is decreasing, the length of the interval is increasing.

2. SE($\bar{x}$) = 0.04;

   Confidence interval with 5% level of significance: (49.6342, 49.7658)
3. SE($\bar{x}$) = 5;

   Confidence interval with 5% level of significance: (9990.2, 10009.8)
4. $\bar{x} = 19.2$, SD = 4.02216, SE($\bar{x}$) = 1.271919

   Confidence interval with 1% level of significance: (15.91845, 22.48155)

   NOTE: $s^2 = \{\sum (x - \bar{x})^2\}/\{n{-}1\}$.
5. $p = 0.42$, SE(p) = 0.047

   Confidence interval with 10% level of significance: (0.3421, 0.4979)

   The same can be expressed in percentage: 34.21% to 49.79%.
6. SE($\bar{x}$) = 0.015
7. n ≥ 384.16; at least 385
8. SE($\bar{x}$) = 0.1061;

   Confidence interval with 5% level of significance: (8.5921, 9.00791)
9. SE($\bar{x}$) = 0.927;

   Confidence interval with 5% level of significance: (13.1831, 16.817)
10. n ≥ 28 18. $n \geq 16$
11. Estimates of P_1 & P_2 are 0.1 and 0.28, respectively. $n \geq 162$
12. Assume $P = 0.05$, $n \geq 385$
13. SE($\bar{x}$) = 0.04971;

    Confidence interval with 1% level of significance: (0.4525, 0.6475)
14. SE($\bar{x}$) = 1.0124;

    Confidence interval with 5% level of significance: (0, 3.9843)
15. SE($\bar{x}$) = 0.0588;

    Confidence interval with 5% level of significance: (0.14, 0.26)
16. SE($\bar{x}$) = 0.0316;

    Confidence interval with 5% level of significance: (0.01684, 0.2316)

Exercise 13

Problem Number	Nature	Ho	H1	Std. Error	Zc	Z_t	Decision
1	Specified mean left-tail	$\mu \geq 200$	$\mu < 200$	1.414	–1.4144	–1.645 $\alpha = 0.05$	Accept Ho
2	Difference of two proportions 2-tail	P1 = P2	P1 ≠ P2	0.0198 $p = 0.2786$	2.529	1.96 $\alpha = 0.05$	Reject Ho
3	Difference of two means 2-tail	$\mu1 = \mu2$	$\mu1 \neq \mu2$	1.561	0	1.96 $\alpha = 0.05$ 2.58 $\alpha = 0.01$	Accept Ho Accept Ho
4	Difference of two means 2-tail	$\mu1 = \mu2$	$\mu1 \neq \mu2$	34.0037	8.8226	1.96 $\alpha = 0.05$	Reject Ho
5	Difference of two proportions 2-tail	P1 = P2	P1 ≠ P2	0.043 $p = 0.29$	1.63	1.96 $\alpha = 0.05$	Accept Ho
6	Difference of two means 2-tail	$\mu1 = \mu2$	$\mu1 \neq \mu2$	6.478	2.316	1.96 $\alpha = 0.05$ 2.58 $\alpha = 0.01$	Reject Ho Accept Ho
7	Specified mean 2-tail	$\mu = 1600$	$\mu \neq 1600$	12	2.5	1.96 $\alpha = 0.05$	Reject Ho
8	Difference of two proportions 2-tail	P1 = P2	P1 ≠ P2	0.0237 $p = 0.47$	–2.11 $\lvert z \rvert = 2.11$	2.58 $\alpha = 0.01$	Accept Ho
9	Difference of two proportions 2-tail	P1 = P2	P1 ≠ P2	0.065 $p = 0.3$	1.54	1.96 $\alpha = 0.05$	Accept Ho
10	Difference of two means 2-tail	$\mu1 = \mu2$	$\mu1 \neq \mu2$	3.40	8.8226	1.96 $\alpha = 0.05$	Reject Ho

(Continued)

Problem Number	Nature	Ho	H1	Std. Error	Zc	Z_t	Decision
11	Difference of two means Right tail	$\mu1 = \mu2$	$\mu1 > \mu2$	1.347	3.713	1.645 $\alpha = 0.05$	Reject Ho
12	Specified proportion 2-tail	$P \leq 0.3$	$P > 0.3$		2.18	1.96 $\alpha = 0.05$	Reject Ho
13	Specified proportion 2-tail	$P = 0.5$	$P \neq 0.5$		2.5298	2.58 $\alpha = 0.01$	Accept Ho
14	Difference of two means 2-tail	$\mu1 = \mu2$	$\mu1 \neq \mu2$	1.35	3.70	1.96 $\alpha = 0.05$	Reject Ho
15	Difference of two proportions Right tail	$P1-P2 = 0.08$	$P1-P2 \neq 0.08$	$p = 0.2$	1.0206	1.96 $\alpha = 0.05$	Accept Ho
16	Small sample Difference of two means 2-tail	$\mu1 = \mu2$	$\mu1 \neq \mu2$	0.63 $s = 1.8787$	2.381	2.75 $\alpha = 0.01$ $df = 34$ t[34,0.005]	Accept Ho
17	Small sample Specified mean	$\mu = 0.5$	$\mu \neq 0.5$	0.0036	2.7634	2.20 $df = 11$ $\alpha = 0.05$	Reject Ho
18	Small sample Difference of two means Right tail	$\mu1 \leq \mu2$	$\mu1 > \mu2$	2.1026 $s = 6.649$	2.756	1.68 $\alpha = 0.05$ $df = 38$ t[38,0.05]	Reject Ho
19	Small sample Difference of two means 2-tail	$\mu1 = \mu2$	$\mu1 \neq \mu2$	20.9862 $s = 40.5491$	5.2416	2.16 $\alpha = 0.05$ $df = 13$ t[13,0.005]	Reject Ho
20	Small sample Difference of two means 2-tail	$\mu1 = \mu2$	$\mu1 \neq \mu2$				Reject Ho

(*Continued*)

Problem Number	Nature	Ho	H1	Std. Error	Zc	Z_t	Decision
21	Small sample Specified mean Right tail	μ = 140	μ > 140		1.71	2.19 *df* = 25 α = 0.05	Reject Ho
22	Small sample Paired *t*-test Left tail	μ1 = μ2 no difference	μ1 < μ2			2.20 *df* = 11 α = 0.05	Reject Ho
23	Chi-square	Independent	Dependent		7.353	3.84 *df* = 1 α = 0.05	Reject Ho
24	Chi-square	Independent	Dependent		19.22	3.84 *df* = 1 α = 0.05	Reject Ho
25	Chi-square	Uniformly distributed	Not uniformly distributed		2.14	11.07 *df* = 5 α = 0.05	Accept Ho
26	Specified mean 2-tail	μ = 150	μ ≠ 150		8.25	3 α = 0	Reject Ho
27	Small sample Paired *t*-test Left tail	μ1 = μ2 no difference	μ1 ≠ μ2		1.4924	2.31 *df* = 8 α = 0.05	Accept Ho
28	Small sample *t*-test diff of mean	μ1 = μ2 no difference	μ1 ≠ μ2		0.9156	2.1 *df* = 19 α = 0.05	Accept Ho
29	Chi-square	The treatments are independent	The treatments are dependent		18.18	3.84 *df* = 1 α = 0.05	Reject Ho
30	Chi-square	The health and social status are independent	The health and social status are dependent		122.44	5.991 *df* = 2 α = 0.05	Reject Ho
31	Specified variance	$\sigma^2 \leq 196$	$\sigma^2 > 196$		35.38	36.415 *df* = 24 α = 0.05	Accept Ho

(*Continued*)

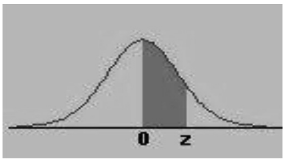
Problem Number	Nature	Ho	H1	Std. Error	Zc	Z_t	Decision
32	Specified variance	$\sigma^2 \leq 200$	$\sigma^2 > 200$		23.75	30.14 $df = 19$ $\alpha = 0.05$	Accept Ho
33	Difference of two means 2-tail	$\mu 1 = \mu 2$	$\mu 1 \neq \mu 2$	175.89	2.843	1.96 $\alpha = 0.05$	Reject Ho
34	F-test	$\sigma_1{}^2 = \sigma_2{}^2$	$\sigma_1{}^2 \neq \sigma_2{}^2$		F = 2.296	2.343 $df = [15,19]$ $\alpha = 0.05$	Accept Ho
35	Chi-square	No difference in opinion	Difference in opinion		1.50	5.991 $df = 2$ $\alpha = 0.05$	Accept Ho
36	Chi-square	There is no relationship between the car and the make	There is a relationship between the car and the make		10.85	12.59 $df = 6$ $\alpha = 0.05$	Accept Ho
37	Chi-square	No relations between the incidence of cavities and brand used [independent]	There is a relationship between the incidence of cavities and brand used		1.91	12.59 $df = 6$ $\alpha = 0.05$	Accept Ho
38	Chi-square	No relations between the opinions about the future of business and their occupation [independent]	There is a relationship between the opinions about the future of business and their occupation		16.68	21.03 $df = 12$ $\alpha = 0.05$	Accept Ho
39	Chi-square	Opinion and the faculty groups are independent	Opinion and the faculty groups are dependent		13.7	9.49 $df = 4$ $\alpha = 0.05$	Reject Ho

(Continued)

Problem Number	Nature	Ho	H1	Std. Error	Zc	Z_t	Decision
40					3.704	4.26 $df = [2,9]$ $\alpha = 0.05$	The difference is not significant
41					4.088	3.49 $df = [3,12]$ $\alpha = 0.05$	There is a significant difference
42					4.52		Ho is rejected
43					2.04 0.875	9.78 [3,6] $\alpha = 0.01$ 10.92 [2,6] $\alpha = 0.01$	Ho is accepted Ho is accepted
44					1.15 11.24	4.757 [3,6] $\alpha = 0.05$ 19.33 [6,2] $\alpha = 0.01$	Ho is accepted Ho is accepted
45					5.52 1.63	3.86 [3,9] $\alpha = 0.05$ 3.86 [3,9] $\alpha = 0.05$	Ho is rejected Ho is accepted

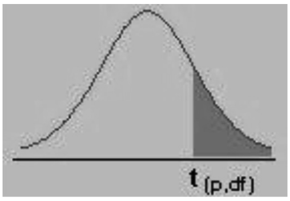
Appendix B: ST Statistical Tables

Standard normal value area between 0 and z



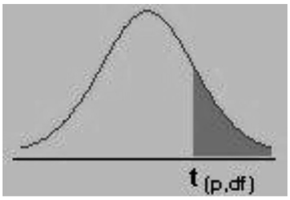
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
+0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	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

t table with right-tail probabilities



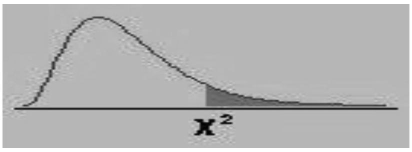
df/p	**0.40**	**0.25**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0005**
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728

t table with right-tail probabilities



df/p	**0.40**	**0.25**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0005**
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
inf	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905

Right-tail areas for the *Chi-square* distribution



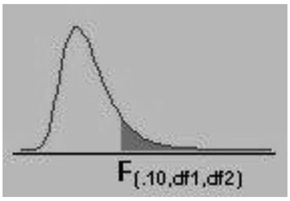
df/area	0.995	0.990	0.975	0.950	0.900	0.750	0.500
1	0.00004	0.00016	0.00098	0.00393	0.01579	0.10153	0.45494
2	0.01003	0.02010	0.05064	0.10259	0.21072	0.57536	1.38629
3	0.07172	0.11483	0.21580	0.35185	0.58437	1.21253	2.36597
4	0.20699	0.29711	0.48442	0.71072	1.06362	1.92256	3.35669
5	0.41174	0.55430	0.83121	1.14548	1.61031	2.67460	4.35146
6	0.67573	0.87209	1.23734	1.63538	2.20413	3.45460	5.34812
7	0.98926	1.23904	1.68987	2.16735	2.83311	4.25485	6.34581
8	1.34441	1.64650	2.17973	2.73264	3.48954	5.07064	7.34412
9	1.73493	2.08790	2.70039	3.32511	4.16816	5.89883	8.34283
10	2.15586	2.55821	3.24697	3.94030	4.86518	6.73720	9.34182
11	2.60322	3.05348	3.81575	4.57481	5.57778	7.58414	10.34100
12	3.07382	3.57057	4.40379	5.22603	6.30380	8.43842	11.34032
13	3.56503	4.10692	5.00875	5.89186	7.04150	9.29907	12.33976
14	4.07467	4.66043	5.62873	6.57063	7.78953	10.16531	13.33927
15	4.60092	5.22935	6.26214	7.26094	8.54676	11.03654	14.33886
16	5.14221	5.81221	6.90766	7.96165	9.31224	11.91222	15.33850
17	5.69722	6.40776	7.56419	8.67176	10.08519	12.79193	16.33818
18	6.26480	7.01491	8.23075	9.39046	10.86494	13.67529	17.33790
19	6.84397	7.63273	8.90652	10.11701	11.65091	14.56200	18.33765
20	7.43384	8.26040	9.59078	10.85081	12.44261	15.45177	19.33743
21	8.03365	8.89720	10.28290	11.59131	13.23960	16.34438	20.33723
22	8.64272	9.54249	10.98232	12.33801	14.04149	17.23962	21.33704
23	9.26042	10.19572	11.68855	13.09051	14.84796	18.13730	22.33688
24	9.88623	10.85636	12.40115	13.84843	15.65868	19.03725	23.33673
25	10.51965	11.52398	13.11972	14.61141	16.47341	19.93934	24.33659
26	11.16024	12.19815	13.84390	15.37916	17.29188	20.84343	25.33646
27	11.80759	12.87850	14.57338	16.15140	18.11390	21.74940	26.33634
28	12.46134	13.56471	15.30786	16.92788	18.93924	22.65716	27.33623
29	13.12115	14.25645	16.04707	17.70837	19.76774	23.56659	28.33613
30	13.78672	14.95346	16.79077	18.49266	20.59923	24.47761	29.33603

df/area	0.250	0.100	0.050	0.025	0.010	0.005
1	1.32330	2.70554	3.84146	5.02389	6.63490	7.87944
2	2.77259	4.60517	5.99146	7.37776	9.21034	10.59663
3	4.10834	6.25139	7.81473	9.34840	11.34487	12.83816
4	5.38527	7.77944	9.48773	11.14329	13.27670	14.86026
5	6.62568	9.23636	11.07050	12.83250	15.08627	16.74960
6	7.84080	10.64464	12.59159	14.44938	16.81189	18.54758
7	9.03715	12.01704	14.06714	16.01276	18.47531	20.27774

(*Continued*)

df/area	0.250	0.100	0.050	0.025	0.010	0.005
8	10.21885	13.36157	15.50731	17.53455	20.09024	21.95495
9	11.38875	14.68366	16.91898	19.02277	21.66599	23.58935
10	12.54886	15.98718	18.30704	20.48318	23.20925	25.18818
11	13.70069	17.27501	19.67514	21.92005	24.72497	26.75685
12	14.84540	18.54935	21.02607	23.33666	26.21697	28.29952
13	15.98391	19.81193	22.36203	24.73560	27.68825	29.81947
14	17.11693	21.06414	23.68479	26.11895	29.14124	31.31935
15	18.24509	22.30713	24.99579	27.48839	30.57791	32.80132
16	19.36886	23.54183	26.29623	28.84535	31.99993	34.26719
17	20.48868	24.76904	27.58711	30.19101	33.40866	35.71847
18	21.60489	25.98942	28.86930	31.52638	34.80531	37.15645
19	22.71781	27.20357	30.14353	32.85233	36.19087	38.58226
20	23.82769	28.41198	31.41043	34.16961	37.56623	39.99685
21	24.93478	29.61509	32.67057	35.47888	38.93217	41.40106
22	26.03927	30.81328	33.92444	36.78071	40.28936	42.79565
23	27.14134	32.00690	35.17246	38.07563	41.63840	44.18128
24	28.24115	33.19624	36.41503	39.36408	42.97982	45.55851
25	29.33885	34.38159	37.65248	40.64647	44.31410	46.92789
26	30.43457	35.56317	38.88514	41.92317	45.64168	48.28988
27	31.52841	36.74122	40.11327	43.19451	46.96294	49.64492
28	32.62049	37.91592	41.33714	44.46079	48.27824	50.99338
29	33.71091	39.08747	42.55697	45.72229	49.58788	52.33562
30	34.79974	40.25602	43.77297	46.97924	50.89218	53.67196

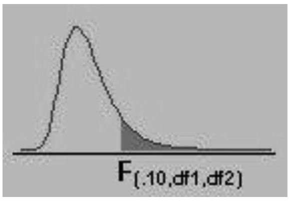
F table for alpha = 0.10.



df_1 / df_2	1	2	3	4	5	6	7	8	9
1	39.8634	49.5000	53.5932	55.8329	57.2401	58.2044	58.9059	59.4389	59.8575
2	8.52632	9.00000	9.16179	9.24342	9.29263	9.32553	9.34908	9.36677	9.38054
3	5.53832	5.46238	5.39077	5.34264	5.30916	5.28473	5.26619	5.25167	5.24000
4	4.54477	4.32456	4.19086	4.10725	4.05058	4.00975	3.97897	3.95494	3.93567
5	4.06042	3.77972	3.61948	3.52020	3.45298	3.40451	3.36790	3.33928	3.31628
6	3.77595	3.46330	3.28876	3.18076	3.10751	3.05455	3.01446	2.98304	2.95774
7	3.58943	3.25744	3.07407	2.96053	2.88334	2.82739	2.78493	2.75158	2.72468
8	3.45792	3.11312	2.92380	2.80643	2.72645	2.66833	2.62413	2.58935	2.56124
9	3.36030	3.00645	2.81286	2.69268	2.61061	2.55086	2.50531	2.46941	2.44034
10	3.28502	2.92447	2.72767	2.60534	2.52164	2.46058	2.41397	2.37715	2.34731
11	3.22520	2.85951	2.66023	2.53619	2.45118	2.38907	2.34157	2.30400	2.27350
12	3.17655	2.80680	2.60552	2.48010	2.39402	2.33102	2.28278	2.24457	2.21352
13	3.13621	2.76317	2.56027	2.43371	2.34672	2.28298	2.23410	2.19535	2.16382

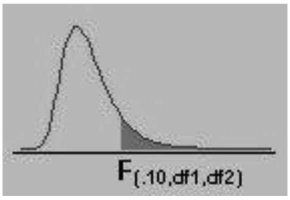
(*Continued*)

F table for alpha = 0.10.



df_1 / df_2	1	2	3	4	5	6	7	8	9
14	3.10221	2.72647	2.52222	2.39469	2.30694	2.24256	2.19313	2.15390	2.12195
15	3.07319	2.69517	2.48979	2.36143	2.27302	2.20808	2.15818	2.11853	2.08621
16	3.04811	2.66817	2.46181	2.33274	2.24376	2.17833	2.12800	2.08798	2.05533
17	3.02623	2.64464	2.43743	2.30775	2.21825	2.15239	2.10169	2.06134	2.02839
18	3.00698	2.62395	2.41601	2.28577	2.19583	2.12958	2.07854	2.03789	2.00467
19	2.98990	2.60561	2.39702	2.26630	2.17596	2.10936	2.05802	2.01710	1.98364
20	2.97465	2.58925	2.38009	2.24893	2.15823	2.09132	2.03970	1.99853	1.96485
21	2.96096	2.57457	2.36489	2.23334	2.14231	2.07512	2.02325	1.98186	1.94797
22	2.94858	2.56131	2.35117	2.21927	2.12794	2.06050	2.00840	1.96680	1.93273
23	2.93736	2.54929	2.33873	2.20651	2.11491	2.04723	1.99492	1.95312	1.91888
24	2.92712	2.53833	2.32739	2.19488	2.10303	2.03513	1.98263	1.94066	1.90625
25	2.91774	2.52831	2.31702	2.18424	2.09216	2.02406	1.97138	1.92925	1.89469
26	2.90913	2.51910	2.30749	2.17447	2.08218	2.01389	1.96104	1.91876	1.88407
27	2.90119	2.51061	2.29871	2.16546	2.07298	2.00452	1.95151	1.90909	1.87427
28	2.89385	2.50276	2.29060	2.15714	2.06447	1.99585	1.94270	1.90014	1.86520
29	2.88703	2.49548	2.28307	2.14941	2.05658	1.98781	1.93452	1.89184	1.85679
30	2.88069	2.48872	2.27607	2.14223	2.04925	1.98033	1.92692	1.88412	1.84896
40	2.83535	2.44037	2.22609	2.09095	1.99682	1.92688	1.87252	1.82886	1.79290
60	2.79107	2.39325	2.17741	2.04099	1.94571	1.87472	1.81939	1.77483	1.73802
120	2.74781	2.34734	2.12999	1.99230	1.89587	1.82381	1.76748	1.72196	1.68425

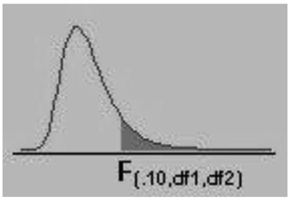
F table for alpha = 0.10



df_1 / df_2	10	12	15	20	24	30	40	60	120
1	60.19498	60.70521	61.22034	61.74029	62.00205	62.26497	62.52905	62.79428	63.06064
2	9.39157	9.40813	9.42471	9.44131	9.44962	9.45793	9.46624	9.47456	9.48289
3	5.23041	5.21562	5.20031	5.18448	5.17636	5.16811	5.15972	5.15119	5.14251
4	3.91988	3.89553	3.87036	3.84434	3.83099	3.81742	3.80361	3.78957	3.77527
5	3.29740	3.26824	3.23801	3.20665	3.19052	3.17408	3.15732	3.14023	3.12279
6	2.93693	2.90472	2.87122	2.83634	2.81834	2.79996	2.78117	2.76195	2.74229
7	2.70251	2.66811	2.63223	2.59473	2.57533	2.55546	2.53510	2.51422	2.49279

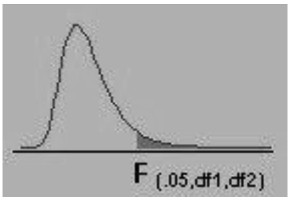
(*Continued*)

F table for alpha = 0.10



df_1 df_2	10	12	15	20	24	30	40	60	120
8	2.53804	2.50196	2.46422	2.42464	2.40410	2.38302	2.36136	2.33910	2.31618
9	2.41632	2.37888	2.33962	2.29832	2.27683	2.25472	2.23196	2.20849	2.18427
10	2.32260	2.28405	2.24351	2.20074	2.17843	2.15543	2.13169	2.10716	2.08176
11	2.24823	2.20873	2.16709	2.12305	2.10001	2.07621	2.05161	2.02612	1.99965
12	2.18776	2.14744	2.10485	2.05968	2.03599	2.01149	1.98610	1.95973	1.93228
13	2.13763	2.09659	2.05316	2.00698	1.98272	1.95757	1.93147	1.90429	1.87591
14	2.09540	2.05371	2.00953	1.96245	1.93766	1.91193	1.88516	1.85723	1.82800
15	2.05932	2.01707	1.97222	1.92431	1.89904	1.87277	1.84539	1.81676	1.78672
16	2.02815	1.98539	1.93992	1.89127	1.86556	1.83879	1.81084	1.78156	1.75075
17	2.00094	1.95772	1.91169	1.86236	1.83624	1.80901	1.78053	1.75063	1.71909
18	1.97698	1.93334	1.88681	1.83685	1.81035	1.78269	1.75371	1.72322	1.69099
19	1.95573	1.91170	1.86471	1.81416	1.78731	1.75924	1.72979	1.69876	1.66587
20	1.93674	1.89236	1.84494	1.79384	1.76667	1.73822	1.70833	1.67678	1.64326
21	1.91967	1.87497	1.82715	1.77555	1.74807	1.71927	1.68896	1.65691	1.62278
22	1.90425	1.85925	1.81106	1.75899	1.73122	1.70208	1.67138	1.63885	1.60415
23	1.89025	1.84497	1.79643	1.74392	1.71588	1.68643	1.65535	1.62237	1.58711
24	1.87748	1.83194	1.78308	1.73015	1.70185	1.67210	1.64067	1.60726	1.57146
25	1.86578	1.82000	1.77083	1.71752	1.68898	1.65895	1.62718	1.59335	1.55703
26	1.85503	1.80902	1.75957	1.70589	1.67712	1.64682	1.61472	1.58050	1.54368
27	1.84511	1.79889	1.74917	1.69514	1.66616	1.63560	1.60320	1.56859	1.53129
28	1.83593	1.78951	1.73954	1.68519	1.65600	1.62519	1.59250	1.55753	1.51976
29	1.82741	1.78081	1.73060	1.67593	1.64655	1.61551	1.58253	1.54721	1.50899
30	1.81949	1.77270	1.72227	1.66731	1.63774	1.60648	1.57323	1.53757	1.49891
40	1.76269	1.71456	1.66241	1.60515	1.57411	1.54108	1.50562	1.46716	1.42476
60	1.70701	1.65743	1.60337	1.54349	1.51072	1.47554	1.43734	1.39520	1.34757
120	1.65238	1.60120	1.54500	1.48207	1.44723	1.40938	1.36760	1.32034	1.26457

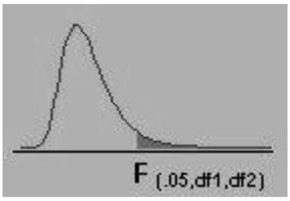
F table for alpha = 0.05.



df_1 df_2	1	2	3	4	5	6	7	8	9
1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988

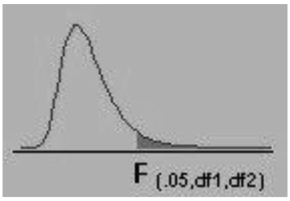
(*Continued*)

F table for alpha = 0.05.



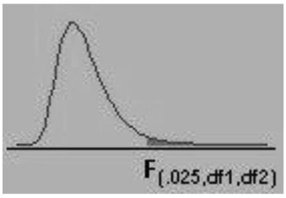
df_1 / df_2	1	2	3	4	5	6	7	8	9
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588

F table for alpha = 0.05.



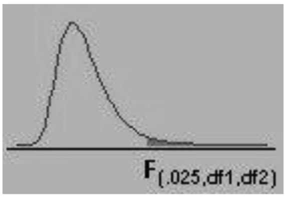
df_1 df_2	10	12	15	20	24	30	40	60	120
1	241.8817	243.9060	245.9499	248.0131	249.0518	250.0951	251.1432	252.1957	253.2529
2	19.3959	19.4125	19.4291	19.4458	19.4541	19.4624	19.4707	19.4791	19.4874
3	8.7855	8.7446	8.7029	8.6602	8.6385	8.6166	8.5944	8.5720	8.5494
4	5.9644	5.9117	5.8578	5.8025	5.7744	5.7459	5.7170	5.6877	5.6581
5	4.7351	4.6777	4.6188	4.5581	4.5272	4.4957	4.4638	4.4314	4.3985
6	4.0600	3.9999	3.9381	3.8742	3.8415	3.8082	3.7743	3.7398	3.7047
7	3.6365	3.5747	3.5107	3.4445	3.4105	3.3758	3.3404	3.3043	3.2674
8	3.3472	3.2839	3.2184	3.1503	3.1152	3.0794	3.0428	3.0053	2.9669
9	3.1373	3.0729	3.0061	2.9365	2.9005	2.8637	2.8259	2.7872	2.7475
10	2.9782	2.9130	2.8450	2.7740	2.7372	2.6996	2.6609	2.6211	2.5801
11	2.8536	2.7876	2.7186	2.6464	2.6090	2.5705	2.5309	2.4901	2.4480
12	2.7534	2.6866	2.6169	2.5436	2.5055	2.4663	2.4259	2.3842	2.3410
13	2.6710	2.6037	2.5331	2.4589	2.4202	2.3803	2.3392	2.2966	2.2524
14	2.6022	2.5342	2.4630	2.3879	2.3487	2.3082	2.2664	2.2229	2.1778
15	2.5437	2.4753	2.4034	2.3275	2.2878	2.2468	2.2043	2.1601	2.1141
16	2.4935	2.4247	2.3522	2.2756	2.2354	2.1938	2.1507	2.1058	2.0589
17	2.4499	2.3807	2.3077	2.2304	2.1898	2.1477	2.1040	2.0584	2.0107
18	2.4117	2.3421	2.2686	2.1906	2.1497	2.1071	2.0629	2.0166	1.9681
19	2.3779	2.3080	2.2341	2.1555	2.1141	2.0712	2.0264	1.9795	1.9302
20	2.3479	2.2776	2.2033	2.1242	2.0825	2.0391	1.9938	1.9464	1.8963
21	2.3210	2.2504	2.1757	2.0960	2.0540	2.0102	1.9645	1.9165	1.8657
22	2.2967	2.2258	2.1508	2.0707	2.0283	1.9842	1.9380	1.8894	1.8380
23	2.2747	2.2036	2.1282	2.0476	2.0050	1.9605	1.9139	1.8648	1.8128
24	2.2547	2.1834	2.1077	2.0267	1.9838	1.9390	1.8920	1.8424	1.7896
25	2.2365	2.1649	2.0889	2.0075	1.9643	1.9192	1.8718	1.8217	1.7684
26	2.2197	2.1479	2.0716	1.9898	1.9464	1.9010	1.8533	1.8027	1.7488
27	2.2043	2.1323	2.0558	1.9736	1.9299	1.8842	1.8361	1.7851	1.7306
28	2.1900	2.1179	2.0411	1.9586	1.9147	1.8687	1.8203	1.7689	1.7138
29	2.1768	2.1045	2.0275	1.9446	1.9005	1.8543	1.8055	1.7537	1.6981
30	2.1646	2.0921	2.0148	1.9317	1.8874	1.8409	1.7918	1.7396	1.6835
40	2.0772	2.0035	1.9245	1.8389	1.7929	1.7444	1.6928	1.6373	1.5766
60	1.9926	1.9174	1.8364	1.7480	1.7001	1.6491	1.5943	1.5343	1.4673
120	1.9105	1.8337	1.7505	1.6587	1.6084	1.5543	1.4952	1.4290	1.3519

F table for alpha = 0.025



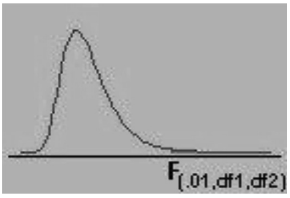
df_1 / df_2	1	2	3	4	5	6	7	8	9
1	647.7890	799.5000	864.1630	899.5833	921.8479	937.1111	948.2169	956.6562	963.2846
2	38.5063	39.0000	39.1655	39.2484	39.2982	39.3315	39.3552	39.3730	39.3869
3	17.4434	16.0441	15.4392	15.1010	14.8848	14.7347	14.6244	14.5399	14.4731
4	12.2179	10.6491	9.9792	9.6045	9.3645	9.1973	9.0741	8.9796	8.9047
5	10.0070	8.4336	7.7636	7.3879	7.1464	6.9777	6.8531	6.7572	6.6811
6	8.8131	7.2599	6.5988	6.2272	5.9876	5.8198	5.6955	5.5996	5.5234
7	8.0727	6.5415	5.8898	5.5226	5.2852	5.1186	4.9949	4.8993	4.8232
8	7.5709	6.0595	5.4160	5.0526	4.8173	4.6517	4.5286	4.4333	4.3572
9	7.2093	5.7147	5.0781	4.7181	4.4844	4.3197	4.1970	4.1020	4.0260
10	6.9367	5.4564	4.8256	4.4683	4.2361	4.0721	3.9498	3.8549	3.7790
11	6.7241	5.2559	4.6300	4.2751	4.0440	3.8807	3.7586	3.6638	3.5879
12	6.5538	5.0959	4.4742	4.1212	3.8911	3.7283	3.6065	3.5118	3.4358
13	6.4143	4.9653	4.3472	3.9959	3.7667	3.6043	3.4827	3.3880	3.3120
14	6.2979	4.8567	4.2417	3.8919	3.6634	3.5014	3.3799	3.2853	3.2093
15	6.1995	4.7650	4.1528	3.8043	3.5764	3.4147	3.2934	3.1987	3.1227
16	6.1151	4.6867	4.0768	3.7294	3.5021	3.3406	3.2194	3.1248	3.0488
17	6.0420	4.6189	4.0112	3.6648	3.4379	3.2767	3.1556	3.0610	2.9849
18	5.9781	4.5597	3.9539	3.6083	3.3820	3.2209	3.0999	3.0053	2.9291
19	5.9216	4.5075	3.9034	3.5587	3.3327	3.1718	3.0509	2.9563	2.8801
20	5.8715	4.4613	3.8587	3.5147	3.2891	3.1283	3.0074	2.9128	2.8365
21	5.8266	4.4199	3.8188	3.4754	3.2501	3.0895	2.9686	2.8740	2.7977
22	5.7863	4.3828	3.7829	3.4401	3.2151	3.0546	2.9338	2.8392	2.7628
23	5.7498	4.3492	3.7505	3.4083	3.1835	3.0232	2.9023	2.8077	2.7313
24	5.7166	4.3187	3.7211	3.3794	3.1548	2.9946	2.8738	2.7791	2.7027
25	5.6864	4.2909	3.6943	3.3530	3.1287	2.9685	2.8478	2.7531	2.6766
26	5.6586	4.2655	3.6697	3.3289	3.1048	2.9447	2.8240	2.7293	2.6528
27	5.6331	4.2421	3.6472	3.3067	3.0828	2.9228	2.8021	2.7074	2.6309
28	5.6096	4.2205	3.6264	3.2863	3.0626	2.9027	2.7820	2.6872	2.6106
29	5.5878	4.2006	3.6072	3.2674	3.0438	2.8840	2.7633	2.6686	2.5919
30	5.5675	4.1821	3.5894	3.2499	3.0265	2.8667	2.7460	2.6513	2.5746
40	5.4239	4.0510	3.4633	3.1261	2.9037	2.7444	2.6238	2.5289	2.4519
60	5.2856	3.9253	3.3425	3.0077	2.7863	2.6274	2.5068	2.4117	2.3344
120	5.1523	3.8046	3.2269	2.8943	2.6740	2.5154	2.3948	2.2994	2.2217

F table for alpha = 0.025.



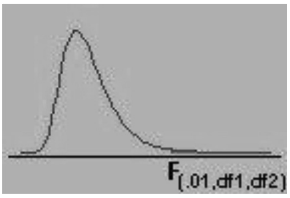
df_1 df_2	10	12	15	20	24	30	40	60	120
1	968.6274	976.7079	984.8668	993.1028	997.2492	1001.414	1005.598	1009.800	1014.020
2	39.3980	39.4146	39.4313	39.4479	39.4562	39.465	39.473	39.481	39.490
3	14.4189	14.3366	14.2527	14.1674	14.1241	14.081	14.037	13.992	13.947
4	8.8439	8.7512	8.6565	8.5599	8.5109	8.461	8.411	8.360	8.309
5	6.6192	6.5245	6.4277	6.3286	6.2780	6.227	6.175	6.123	6.069
6	5.4613	5.3662	5.2687	5.1684	5.1172	5.065	5.012	4.959	4.904
7	4.7611	4.6658	4.5678	4.4667	4.4150	4.362	4.309	4.254	4.199
8	4.2951	4.1997	4.1012	3.9995	3.9472	3.894	3.840	3.784	3.728
9	3.9639	3.8682	3.7694	3.6669	3.6142	3.560	3.505	3.449	3.392
10	3.7168	3.6209	3.5217	3.4185	3.3654	3.311	3.255	3.198	3.140
11	3.5257	3.4296	3.3299	3.2261	3.1725	3.118	3.061	3.004	2.944
12	3.3736	3.2773	3.1772	3.0728	3.0187	2.963	2.906	2.848	2.787
13	3.2497	3.1532	3.0527	2.9477	2.8932	2.837	2.780	2.720	2.659
14	3.1469	3.0502	2.9493	2.8437	2.7888	2.732	2.674	2.614	2.552
15	3.0602	2.9633	2.8621	2.7559	2.7006	2.644	2.585	2.524	2.461
16	2.9862	2.8890	2.7875	2.6808	2.6252	2.568	2.509	2.447	2.383
17	2.9222	2.8249	2.7230	2.6158	2.5598	2.502	2.442	2.380	2.315
18	2.8664	2.7689	2.6667	2.5590	2.5027	2.445	2.384	2.321	2.256
19	2.8172	2.7196	2.6171	2.5089	2.4523	2.394	2.333	2.270	2.203
20	2.7737	2.6758	2.5731	2.4645	2.4076	2.349	2.287	2.223	2.156
21	2.7348	2.6368	2.5338	2.4247	2.3675	2.308	2.246	2.182	2.114
22	2.6998	2.6017	2.4984	2.3890	2.3315	2.272	2.210	2.145	2.076
23	2.6682	2.5699	2.4665	2.3567	2.2989	2.239	2.176	2.111	2.041
24	2.6396	2.5411	2.4374	2.3273	2.2693	2.209	2.146	2.080	2.010
25	2.6135	2.5149	2.4110	2.3005	2.2422	2.182	2.118	2.052	1.981
26	2.5896	2.4908	2.3867	2.2759	2.2174	2.157	2.093	2.026	1.954
27	2.5676	2.4688	2.3644	2.2533	2.1946	2.133	2.069	2.002	1.930
28	2.5473	2.4484	2.3438	2.2324	2.1735	2.112	2.048	1.980	1.907
29	2.5286	2.4295	2.3248	2.2131	2.1540	2.092	2.028	1.959	1.886
30	2.5112	2.4120	2.3072	2.1952	2.1359	2.074	2.009	1.940	1.866
40	2.3882	2.2882	2.1819	2.0677	2.0069	1.943	1.875	1.803	1.724
60	2.2702	2.1692	2.0613	1.9445	1.8817	1.815	1.744	1.667	1.581
120	2.1570	2.0548	1.9450	1.8249	1.7597	1.690	1.614	1.530	1.433

F table for alpha = 0.01.



df_1 / df_2	1	2	3	4	5	6	7	8	9
1	4052.181	4999.500	5403.352	5624.583	5763.650	5858.986	5928.356	5981.070	6022.473
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718
120	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559

F table for alpha = 0.01.



df_1 / df_2	10	12	15	20	24	30	40	60	120
1	6055.847	6106.321	6157.285	6208.730	6234.631	6260.649	6286.782	6313.030	6339.391
2	99.399	99.416	99.433	99.449	99.458	99.466	99.474	99.482	99.491
3	27.229	27.052	26.872	26.690	26.598	26.505	26.411	26.316	26.221
4	14.546	14.374	14.198	14.020	13.929	13.838	13.745	13.652	13.558
5	10.051	9.888	9.722	9.553	9.466	9.379	9.291	9.202	9.112
6	7.874	7.718	7.559	7.396	7.313	7.229	7.143	7.057	6.969
7	6.620	6.469	6.314	6.155	6.074	5.992	5.908	5.824	5.737
8	5.814	5.667	5.515	5.359	5.279	5.198	5.116	5.032	4.946
9	5.257	5.111	4.962	4.808	4.729	4.649	4.567	4.483	4.398
10	4.849	4.706	4.558	4.405	4.327	4.247	4.165	4.082	3.996
11	4.539	4.397	4.251	4.099	4.021	3.941	3.860	3.776	3.690
12	4.296	4.155	4.010	3.858	3.780	3.701	3.619	3.535	3.449
13	4.100	3.960	3.815	3.665	3.587	3.507	3.425	3.341	3.255
14	3.939	3.800	3.656	3.505	3.427	3.348	3.266	3.181	3.094
15	3.805	3.666	3.522	3.372	3.294	3.214	3.132	3.047	2.959
16	3.691	3.553	3.409	3.259	3.181	3.101	3.018	2.933	2.845
17	3.593	3.455	3.312	3.162	3.084	3.003	2.920	2.835	2.746
18	3.508	3.371	3.227	3.077	2.999	2.919	2.835	2.749	2.660
19	3.434	3.297	3.153	3.003	2.925	2.844	2.761	2.674	2.584
20	3.368	3.231	3.088	2.938	2.859	2.778	2.695	2.608	2.517
21	3.310	3.173	3.030	2.880	2.801	2.720	2.636	2.548	2.457
22	3.258	3.121	2.978	2.827	2.749	2.667	2.583	2.495	2.403
23	3.211	3.074	2.931	2.781	2.702	2.620	2.535	2.447	2.354
24	3.168	3.032	2.889	2.738	2.659	2.577	2.492	2.403	2.310
25	3.129	2.993	2.850	2.699	2.620	2.538	2.453	2.364	2.270
26	3.094	2.958	2.815	2.664	2.585	2.503	2.417	2.327	2.233
27	3.062	2.926	2.783	2.632	2.552	2.470	2.384	2.294	2.198
28	3.032	2.896	2.753	2.602	2.522	2.440	2.354	2.263	2.167
29	3.005	2.868	2.726	2.574	2.495	2.412	2.325	2.234	2.138
30	2.979	2.843	2.700	2.549	2.469	2.386	2.299	2.208	2.111
40	2.801	2.665	2.522	2.369	2.288	2.203	2.114	2.019	1.917
60	2.632	2.496	2.352	2.198	2.115	2.028	1.936	1.836	1.726
120	2.472	2.336	2.192	2.035	1.950	1.860	1.763	1.656	1.533

Index

accuracy 12–15, 31, 209, 219
action on results 3–4
addition rule for probability 143
advantages of: arithmetic mean 45–6; diagrammatic representation 31; mode 62; primary data 11, 19; sampling 191; SD 85; weighted mean 71
alternate hypothesis 225
ANOVA 228, 301, 323
application of statistics in business 2
area under normal curve 180–1
area under standard normal curve 181

bar diagram 31–4
Baye's theorem 150
Bernoulli process 169, 171
binomial distribution 169–71, 173, 175
bivariate data 10

census method 16
characteristics of discrete distribution 163
characteristics of good table 23
chi-square distribution 270–1
chi-square test 271, 287, 319
chronological classification 21
classification of data 9
cluster sampling 198
coefficient of quartile deviation 80, 82
coefficient of variation 89–90
comparison of mean, median, and mode 69
completeness 15
component bar chart 31, 33
composite mean 46
compound probability 147
conditional probability 147, 150, 152
conditions of binomial distribution 170
confidence interval estimation 209, 215
consistency 15, 74, 208
construction of frequency distribution 25, 28–30, 39–41
continuous data 23, 44–5, 51
continuous probability distribution 179
continuous random variable 161, 179
continuous variable 23
convenience sampling 197
correlation 107–9, 115, 118, 120
cumulative frequency 29

data 9; analysis 4; collection 4; presentation 21; sources 9, 11; structures 9
deciles 58
degrees of freedom 191
descriptive statistics 2
determining sample size 219
deviational bar diagram 33–4
diagrammatic representation 31
different methods of sampling 1, 16
disadvantages of primary data 11, 19
discrete data 23, 44, 51, 61
discrete random variable 161
discrete variable 23
dispersion 79, 89, 97
distrust of statistics 6

editing primary data 15
equally like events 140
errors due to bias 15
estimate 1, 7, 10, 206–8
estimating error 15
estimator 15, 203, 207–8
event 139–42, 146
exclusive method 26
experiment 139
external source 11

F-distribution 294–5
frequency 25, 28–9, 36, 44–5, 51, 61, 141; curve 31, 36–7; distributions 25, 28–9; polygon 31, 36–7
functions of statistics 5

geographical classification 21
geometric mean 43, 71
graphical method to find median 56
graphical method to find mode 68
group correlation 107–8

histogram 31, 36

inclusive method 26–7
inferential statistics 2
internal source 11
interval estimation 209, 215
interval scale 24

kurtosis 97, 102

law of inertia of large numbers 7, 192
law of statistical regularity 7, 192
leptokurtic 102
levels of measurement 24
limitations of statistics 5, 7
line diagram 31, 38

mail questionnaire method 13
mathematical expectation 165
mathematical properties of arithmetic mean 46
mean: of binomial distribution 170; deviation 82; of Poisson distribution 174
measures of central tendency 43
median 43, 50, 56, 69
mesokurtic 102
methodology of statistical testing 228
mode 43, 61
moments 97, 100
multiple bar chart 31
multiplication rule on probability 146
multistage sampling 196
multivariate data 9–10
mutually exclusive events 140, 144, 150

nature of statistical law 7
need for sampling distribution 202
nominal scale 24
non parametric test 228
non probability sample 192
non-random sampling methods 197
non sampling errors 192
normal distribution 102, 179, 181–2, 202–3
null hypothesis 225
number of classes 225

observation method 12
ogive curve 31, 38, 56–8
one tailed test-left tailed 227
one tailed test-right tailed 226
one-way classification 301, 303–4, 307
open-end method 26
ordinal scale 24
outcome 7, 15, 115, 127, 139–40, 161–2, 167, 169–70, 199

paired t-test 267
parameter 190, 202, 206–9, 225
parametric test 225, 228
percentiles 58
phases of statistical decision making process 3
pie diagram 31, 34
pilot survey 14
platykurtic 102
point estimation 206
Poisson distribution 170, 173–5
population 189; mean 231–2, 234; SD 221, 236; size 191, 193–4, 202, 204, 237
possible errors in secondary data 15
primary sources 11
probability 139, 141–2; distribution 161–2; function 163; sample 192
probable error of coefficient of correlation 118
properties of: best average 43; binomial distribution 170; chi-square distribution 271; correlation 108; F-distribution 295; good point estimators 208; normal distribution 182; Poisson distribution 174; t-distribution 254
purposive sampling 197

qualitative classification 21
quantitative classification 22
quartile deviation (DD) 80
quartiles 58
questionnaire design 14
questionnaire -example:1 17
questionnaire method 12–13
quota sampling 197

random errors 192
random sampling 193
random variable 161
range 79
rank correlation 107–8, 120–1
ratio scale 25
regression equations 125
relative advantages of median 52
relative efficiency 208
relative frequency 29, 140–1
repeated ranks 122
responsibility of decision maker 4

sample 189; design 4, 190; space 139
sampling: distribution based on means of samples 205; distributions 199; errors 192; methods 1, 16
scatter diagram 108
secondary data - usage 16
secondary sources 11
sequential sampling 199
size of classes 26
skewness 97

standard deviation 84
standard error 204
statistical inference 189, 202
statistics 1–3, 190
stratified sampling 195
student's t-distribution 253
study design 3–4
sufficiency 208
survey design 13
systematic sampling 194

tabular form 22
tally mark 28
telephonic interview method 13
test for: difference two populations variances 295; equality of two-populations means large sample 239; equality of two-populations means small sample 260; equality of two population variances 238; equality of two standard deviations 252; goodness of fit 272; independence of attributes 279; population proportion 243; specified mean large sample 229; specified population variance 290
textual form 22
transcribing error 15
two tailed test 226
two way classification 301, 308, 314
type I error 227
type II error 227
types of averages 43
types of class intervals 26
types of variables and data 23

univariate data 9

variance 85, 255, 295; of binomial distribution 171; of Poisson distribution 174; of random variable 166

weighted arithmetic mean 70

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com