

UNIVERSITY MBA SERIES

Statistics for Management

FOR VTU



J. K. Sharma

Statistics for Management

J. K. Sharma

Professor

**Amity Business School
Amity University, Noida**

PEARSON

Chennai • Delhi • Chandigarh

Copyright © 2012 Dorling Kindersley (India) Pvt. Ltd.

Licensees of Pearson Education in South Asia

No part of this eBook may be used or reproduced in any manner whatsoever without the publisher's prior written consent.

This eBook may or may not include all assets that were part of the print version. The publisher reserves the right to remove any material present in this eBook at any time.

ISBN 9788131765029

eISBN 9788131776353

Head Office: A-8(A), Sector 62, Knowledge Boulevard, 7th Floor, NOIDA 201 309, India

Registered Office: 11 Local Shopping Centre, Panchsheel Park, New Delhi 110 017, India

Preface

Statistical thinking enhance our understanding of how life works, allows control over some societal issues and helps individuals make informed decisions. I am sure after studying this book your skills in business decision-making and understanding of the problems of business and industry will improve.

This book has been written as a practical response to the needs of students who want to obtain a reasonable grasp of basic statistical techniques or methods in a limited time. The emphasis throughout the book is on understanding through practice, interpretation of results and their application to the real-life problems. Statistical theory and derivation of formulae are deliberately kept to a minimum. This will encourage students who lack confidence in their mathematical ability to understand statistical techniques.

Each chapter of the book includes the necessary theory and methods of carrying out the various techniques and analysis. A large number of solved examples and self practice problems (all with hints and answers) are provided to motivate students to apply statistical techniques to real data and draw statistical inferences. Other than providing useful guidance to the students in several professional and competitive examinations, this book should serve as core textbook for the students of

- BBA, BCA, BCom
- PGDBM, MBA, MCom, MA (Eco)
- MCA, BE, BTech (Computer Science)
- CA, ICWA, AMIE

I am indebted to all my students, friends and colleagues for their helpful input while writing this book. In particular, I am thankful to Prof V K Bhalla and Prof R P Hooda for their valuable suggestions and encouragement.

In writing this book I have benefited immensely by referring to several books and research papers. I express my gratitude to authors, publishers and institutions of all such books and papers.

I would like to thank the editorial team, in particular Mr. Sanjay Singh and Mr. Raza Khan at Pearson Education, for assistance in bringing out this book. Thanks are also due to Mr. Dinesh Kaushik and Mr. Pawan Tyagi for their cooperation in designing the layout of the book. Finally, I am thankful to my wife and children for their patience, understanding, love and assistance in making this book a reality. It is to them that I dedicate this book.

Suggestions and comments to improve the book in content and style are always welcome and will be greatly appreciated and acknowledged.

This page is intentionally left blank.

Contents

Preface	iii	
CHAPTER 1 STATISTICS: AN OVERVIEW	1-26	
1.1 Reasons For Learning Statistics	1	
1.2 Growth and Development of Statistics	2	
1.3 Statistical Thinking and Analysis	3	
1.4 Statistics Defined	3	
1.5 Types of Statistical Methods	5	
1.6 Importance and Scope of Statistics	6	
1.7 Limitations of Statistics	7	
1.8 How to Lie with Statistics	8	
<i>Conceptual Questions 1A</i>	10	
1.9 Need for Data	11	
1.10 Principles of Measurement	12	
1.11 Sources of Data	16	
<i>Conceptual Questions 1B</i>	25	
CHAPTER 2 DATA CLASSIFICATION, TABULATION AND PRESENTATION	27-80	
2.1 Introduction	27	
2.2 Classification of Data	27	
2.3 Organizing Data Using Data Array	30	
<i>Conceptual Questions 2A</i>	42	
<i>Self-Practice Problems 2A</i>	43	
<i>Hints and Answers</i>	44	
2.4 Tabulation of Data	44	
<i>Conceptual Questions 2B</i>	52	
<i>Self-Practice Problems 2B</i>	52	
<i>Hints and Answers</i>	54	
2.5 Graphical Presentation of Data	55	
2.6 Types of Diagrams	57	
2.7 Exploratory Data Analysis	71	
<i>Conceptual Questions 2C</i>	73	
		<i>Self-Practice Problems 2C</i> 73
		<i>Hints and Answers</i> 75
		<i>Formulae Used</i> 76
		<i>Review Self-Practice Problems</i> 76
		<i>Case Studies</i> 78
		CHAPTER 3 MEASURES OF CENTRAL TENDENCY 81-130
		3.1 Introduction
		3.2 Objectives of Averaging
		3.3 Requisites of a Measure of Central Tendency
		3.4 Measures of Central Tendency
		3.5 Mathematical Averages
		<i>Conceptual Questions 3A</i> 99
		<i>Self-Practice Problems 3A</i> 99
		<i>Hints and Answers</i> 101
		3.6 Geometric Mean
		<i>Conceptual Questions 3B</i> 105
		<i>Self-Practice Problems 3B</i> 105
		<i>Hints and Answers</i> 106
		3.7 Harmonic Mean
		3.8 Relationship among A.M., G.M., and H.M.
		<i>Self-Practice Problems 3C</i> 108
		<i>Hints and Answers</i> 108
		3.9 Averages of Position
		3.10 Partition Values—Quartiles, Deciles, and Percentiles
		<i>Conceptual Questions 3C</i> 116
		<i>Self-Practice Problems 3D</i> 116
		<i>Hints and Answers</i> 117
		3.11 Mode
		3.12 Relationship Between Mean, Median, and Mode

CHAPTER 3 3.13 Comparison Between Measures of Central Tendency 122 <i>Conceptual Questions 3C</i> 122 <i>Self-Practice Problems 3E</i> 123 <i>Hint and Answers</i> 125 <i>Formulae Used</i> 125 <i>Review Self-Practice Problems</i> 125 <i>Hints and Answers</i> 127 <i>Case Studies</i> 129	CHAPTER 6 PROBABILITY DISTRIBUTIONS 205-248 6.1 Introduction 205 6.2 Probability Distribution Function (<i>pdf</i>) 206 6.3 Cumulative Probability Distribution Function (<i>cdf</i>) 207 6.4 Expected Value and Variance of a Random Variable 209 <i>Conceptual Questions 6A</i> 213 <i>Self-Practice Problems 6A</i> 214 <i>Hints and Answers</i> 214 6.5 Discrete Probability Distributions 215 <i>Conceptual Questions 6B</i> 220 <i>Self-Practice Problems 6B</i> 220 <i>Hints and Answers</i> 221 <i>Conceptual Questions 6C</i> 228 <i>Self-Practice Problems 6C</i> 228 <i>Hints and Answers</i> 229 6.6 Continuous Probability Distributions 231 <i>Conceptual Questions 6D</i> 241 <i>Self-Practice Problems 6D</i> 241 <i>Hints and Answers</i> 242 <i>Formulae Used</i> 243 <i>Review Self-Practice Problems</i> 244 <i>Hints and Answers</i> 245
CHAPTER 4 MEASURES OF DISPERSION 131-168 4.1 Introduction 131 4.2 Significance of Measuring Dispersion 132 4.3 Classification of Measures of Dispersion 133 4.4 Distance Measures 134 <i>Conceptual Questions 4A</i> 138 <i>Self-Practice Problems 4A</i> 138 <i>Hints and Answers</i> 139 4.5 Average Deviation Measures 140 <i>Conceptual Questions 4B</i> 157 <i>Self-Practice Problems 4B</i> 158 <i>Hints and Answers</i> 160 <i>Formulae Used</i> 161 <i>Review Self-Practice Problems</i> 161 <i>Hints and Answers</i> 164 <i>Case Studies</i> 166	CHAPTER 7 SAMPLING AND SAMPLING DISTRIBUTIONS 249-274 7.1 Introduction 249 7.2 Reasons of Sample Survey 250 7.3 Types of Bias during Sample Survey 250 7.4 Population Parameters and Sample Statistics 251 7.5 Principles of Sampling 251 7.6 Sampling Methods 252 7.7 Sampling Distributions 257 <i>Conceptual Questions 7A</i> 259 7.8 Sampling Distribution of Sample Mean 260 <i>Self-Practice Problems 7A</i> 268 <i>Hints and Answers</i> 269 7.9 Sampling Distribution of Sample Proportion 270 <i>Self-Practice Problems 7B</i> 272 <i>Hints and Answers</i> 272 <i>Formulae Used</i> 273 <i>Review Self-Practice Problems</i> 273 <i>Hints and Answers</i> 274
CHAPTER 5 FUNDAMENTALS OF PROBABILITY 169-204 5.1 Introduction 169 5.2 Concepts of Probability 169 5.3 Definition of Probability 172 5.4 Counting Rules for Determining the Number of Outcomes 174 <i>Conceptual Questions 5A</i> 177 <i>Self-Practice Problems 5A</i> 177 <i>Hints and Answers</i> 178 5.5 Rules of Probability and Algebra of Events 179 5.6 Probability Tree Diagram 190 <i>Self-Practice Problems 5B</i> 191 <i>Hints and Answers</i> 193 5.7 Bayes' Theorem 195 <i>Self-Practice Problems 5C</i> 197 <i>Hints and Answers</i> 198 <i>Formulae Used</i> 199 <i>Review Self-Practice Problems</i> 199 <i>Hints and Answers</i> 201 <i>Case Studies</i> 203	

CHAPTER 8 HYPOTHESIS TESTING	275–326	CHAPTER 10 CORRELATION ANALYSIS	351–384
8.1 Introduction 275		10.1 Introduction 351	
8.2 Hypothesis and Hypothesis Testing 275		10.2 Significance of Measuring Correlation 352	
8.3 The Rationale for Hypothesis Testing 276		10.3 Correlation and Causation 352	
8.4 General Procedure for Hypothesis Testing 277		10.4 Types of Correlations 353	
8.5 Direction of the Hypothesis Test 280		10.5 Methods of Correlation Analysis 354	
8.6 Errors in Hypothesis Testing 281		<i>Self-Practice Problems 10A</i> 365	
<i>Conceptual Questions 8A</i> 284		<i>Hints and Answers</i> 366	
8.7 Hypothesis Testing for Population Parameters with Large Samples 284		<i>Self-Practice Problems 10B</i> 374	
<i>Self-Practice Problems 8A</i> 293		<i>Hints and Answers</i> 375	
<i>Hints and Answers</i> 294		10.6 Hypothesis Testing for Correlation Coefficient 375	
8.8 Hypothesis Testing for Single Population Proportion 295		<i>Conceptual Questions 10A</i> 379	
8.9 Hypothesis Testing for a Binomial Proportion 298		<i>Self-Practice Problems 10C</i> 380	
<i>Self-Practice Problems 8B</i> 299		<i>Hints and Answers</i> 380	
<i>Hints and Answers</i> 300		<i>Formulae Used</i> 381	
8.10 Hypothesis Testing for Population Mean with Small Samples 301		<i>Review-Self Practice Problems</i> 382	
<i>Self-Practice Problems 8C</i> 312		<i>Hints and Answers</i> 383	
<i>Hints and Answers</i> 313			
8.11 Hypothesis Testing Based on F-Distribution 315			
<i>Self-Practice Problems 8D</i> 318			
<i>Hints and Answers</i> 319			
<i>Formulae Used</i> 320			
<i>Review-Self Practice Problems</i> 321			
<i>Hints and Answers</i> 323			
CHAPTER 9 ANALYSIS OF VARIANCE	327–350	CHAPTER 11 REGRESSION ANALYSIS	385–416
9.1 Introduction 327		11.1 Introduction 385	
9.2 Analysis of Variance Approach 329		11.2 Advantages of Regression Analysis 386	
9.3 Testing Equality of Population (Treatment) Means: One-way Classification 329		11.3 Types of Regression Models 386	
9.4 Inferences About Population (Treatment) Means 337		11.4 Estimation : The Method of Least Squares 388	
<i>Self-Practice Problems 9A</i> 338		11.5 Assumptions for a Simple Linear Regression Model 389	
<i>Hints and Answers</i> 338		11.6 Parameters of Simple Linear Regression Model 389	
9.5 Testing Equality of Population (Treatment) Means: Two-Way Classification 339		11.7 Methods to Determine Regression Coefficients 391	
<i>Conceptual Questions 9A</i> 343		<i>Self-Practice Problems 11A</i> 401	
<i>Self-Practice Problems 9B</i> 344		<i>Hints and Answers</i> 403	
<i>Hints and Answers</i> 345		11.8 Standard Error of Estimate and Prediction Intervals 405	
<i>Formulae Used</i> 346		<i>Conceptual Questions 11A</i> 410	
<i>Review Self-Practice Problems</i> 347		<i>Formulae Used</i> 411	
<i>Hints and Answers</i> 348		<i>Review Self-Practice Problems</i> 411	
<i>Case Studies</i> 349		<i>Hints and Answers</i> 413	
		<i>Case Studies</i> 415	
CHAPTER 12 FORECASTING AND TIME SERIES ANALYSIS		417–464	
12.1 Introduction 417			
12.2 Types of Forecasts 418			
12.3 Timing of Forecasts 418			
12.4 Forecasting Methods 419			
12.5 Steps of Forecasting 421			
12.6 Time Series Analysis 421			
12.7 Time Series Decomposition Models 422			
		<i>Conceptual Questions 12A</i> 423	

<p>12.8 Quantitative Forecasting Methods 424 <i>Self-Practice Problems 12A</i> 434 <i>Hints and Answers</i> 435</p> <p>12.9 Trend Projection Methods 437 <i>Self-Practice Problems 12B</i> 443 <i>Hints and Answers</i> 444</p> <p>12.10 Measurement of Seasonal Effects 445</p> <p>12.11 Measurement of Cyclical Variations Residual Method 457</p> <p>12.12 Measurement of Irregular Variations 457 <i>Conceptual Questions 12B</i> 457 <i>Self-Practice Problems 12C</i> 458 <i>Hints and Answers</i> 459 <i>Formulae Used</i> 460 <i>Review Self-Practice Problems</i> 461 <i>Hints and Answers</i> 462</p>	<p>CHAPTER 14 SKEWNESS, MOMENTS AND KURTOSIS 515-537</p> <p>14.1 Introduction 515 14.2 Measures of Skewness 516 <i>Conceptual Questions 14A</i> 522 <i>Self-Practice Problems 14A</i> 523 <i>Hints and Answers</i> 524</p> <p>14.3 Moments 525 14.4 Kurtosis 530 <i>Conceptual Questions 14B</i> 533 <i>Self-Practice Problems 14B</i> 533 <i>Hints and Answers</i> 534 <i>Formulae Used</i> 534 <i>Review Self-Practice Problems</i> 535 <i>Hints and Answers</i> 536</p>
<p>CHAPTER 13 INDEX NUMBERS 465-514</p> <p>13.1 Introduction 465 13.2 Index Number Defined 466 13.3 Types of Index Numbers 466 13.4 Characteristics and uses of Index Numbers 468 <i>Conceptual Questions 13A</i> 469 13.5 Methods for Construction of Price Indexes 470 13.6 Unweighted Price Indexes 470 <i>Self-Practice Problems 13A</i> 474 <i>Hints and Answers</i> 475</p> <p>13.7 Weighted Price Indexes 477 13.8 Quantity or Volume Indexes 484 13.9 Value Indexes 486 <i>Self-Practice Problems 13B</i> 487 <i>Hints and Answers</i> 488</p> <p>13.10 Tests of Adequacy of Indexes 489 13.11 Chain Indexes 492 13.12 Applications of Index Numbers 496 <i>Self-Practice Problems 13C</i> 500 <i>Hints and Answers</i> 501</p> <p>13.13 Consumer Price Indexes 504 13.14 Problems of Index Number Construction 507 <i>Conceptual Questions 13B</i> 509 <i>Formulae Used</i> 509 <i>Review Self-Practice Problems</i> 510 <i>Hints and Answers</i> 511</p>	<p>CHAPTER 15 CHI-SQUARE AND OTHER NON-PARAMETRIC TESTS 539-575</p> <p>15.1 Introduction 539 15.2 Advantages and Limitations of Non-parametric Methods 540 15.3 The Chi-Square Distribution 540 15.4 The Chi-square Test-statistic 542 15.5 Applications of χ^2 Test 543 <i>Self-Practice Problems 15A</i> 547 <i>Hints and Answers</i> 548</p> <p><i>Self-Practice Problems 15B</i> 554 <i>Hints and Answers</i> 554</p> <p><i>Conceptual Questions 15A</i> 559 15.6 The Sign Test for Paired Data 559 15.7 Runs Test for Randomness 561 15.8 Mann-Whitney U-Test 563 15.9 Wilcoxon Matched Pairs Test 565 15.10 Kruskal-Wallis Test 567 <i>Self-Practice Problems 15C</i> 569 <i>Hints and Answers</i> 571</p> <p><i>Self-Practice Problems</i> 572 <i>Hints and Answers</i> 574</p>
<p>APPENDICES 577-589</p> <p>MODEL QUESTION PAPERS 591-603</p> <p>SOLUTION TO MODEL QUESTION PAPER-I 605-622</p> <p>SOLUTION TO MODEL QUESTION PAPER-II 623-639</p>	

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

—H. G. Wells

Statistics: An Overview

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- present a broad overview of statistics as a subject.
- bring out applications of statistics and its usefulness in managerial decision-making.
- describe the data collection process.
- understand basic concepts of questionnaire design and measurement scales.

1.1 REASONS FOR LEARNING STATISTICS

H. G. Wells' statement that "statistical thinking will one day be as necessary as the ability to read and write" is valid in the context of today's competitive business environment where many organizations find themselves data-rich but information-poor. Thus, for decision-makers, it is important to develop the ability to extract meaningful information from raw data to make better decisions. It is possible only through the careful analysis of data guided by statistical thinking.

The reason for analysis of **data** is an understanding of *variation and its causes* in any phenomenon. Since variation is present in all phenomena, therefore knowledge of it leads to better decisions about a phenomenon that produced the data. It is from this perspective that the learning of statistics enables the decision-maker to understand how to

- present and describe information (data) so as to improve decisions.
- draw conclusions about the large **population** based upon information obtained from samples.
- seek out relationship between pair of variables to improve processes.
- obtain reliable forecasts of statistical variables of interest.

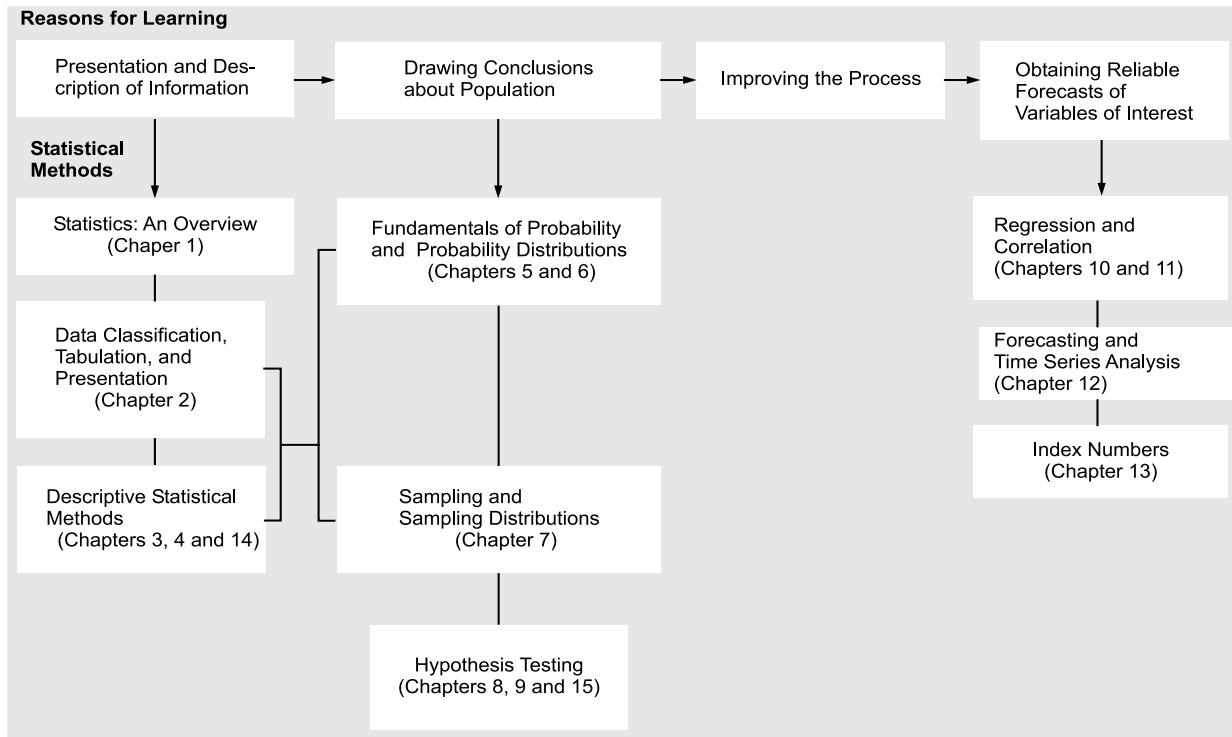
Data: A collection of observations of one or more variables of interest.

Population: A collection of all elements (units or variables) of interest.

Thus a statistical study might be a simple exploration enabling us to gain insight into a virtually unknown situation or it might be a sophisticated analysis to produce numerical confirmation or a reflection of some widely held belief.

As shown in Fig. 1.1, the text matter of the book has been organized keeping in view these four reasons for learning statistics.

Figure 1.1
Flow Chart of Reasons For Learning Statistics



1.2 GROWTH AND DEVELOPMENT OF STATISTICS

Statistics: The art and science of collecting, analysing, presenting, and interpreting data.

The views commonly held about **statistics** are numerous, but often incomplete. It has different meanings to different people depending largely on its use. For example, (i) for a cricket fan, statistics refers to numerical information or data relating to the runs scored by a cricketer; (ii) for an environmentalist, statistics refers to information on the quantity of pollutants released into the atmosphere by all types of vehicles in different cities; (iii) for the census department, statistics consists of information about the birth rate and the sex ratio in different states; (iv) for a share broker, statistics is the information on changes in share prices over a period of time; and so on.

The average person perceives statistics as a column of figures, various types of graphs, tables and charts showing the increase and/or decrease in per capita income, wholesale price index, industrial production, exports, imports, crime rate and so on. The sources of such statistics for a common man are newspapers, magazines/journals, reports/bulletins, radio, and television. In all such cases, the relevant data are collected; numbers manipulated and information presented with the help of figures, charts, diagrams, and pictograms; probabilities are quoted, conclusions reached, and discussions held. Efforts to understand and find a solution (with certain degree of precision) to problems pertaining to social, political, economic, and cultural activities, seem to be unending. All such efforts are guided by the use of methods drawn from the field of statistics.

The development of mathematics in relation to the probability theory and the advent of fast-speed computers have substantially changed the field of statistics in the last few decades. The use of computer software, such as SAS and SPSS, have brought about a technological revolution. The increasing use of spreadsheet packages like Lotus 1-2-3 and Microsoft Excel have led to the incorporation of statistical features in these packages. These softwares have made the task of statistical analysis quite convenient and feasible.

1.3 STATISTICAL THINKING AND ANALYSIS

An integral part of the managerial approach focuses on the quality of products manufactured or services provided by an organization. This approach requires the application of certain statistical methods and the statistical thinking by the management of the organization. *Statistical thinking can be defined as the thought process that focuses on ways to identify, control, and reduce variations present in all phenomena.* A better understanding of a phenomenon through statistical thinking and use of statistical methods for data analysis, enhances opportunities for improvement in the quality of products or services. Statistical thinking also allows one to recognize and make interpretations of the variations in a process.

As shown in Fig. 1.2, *management philosophy* acts as a guide for laying a solid foundation for total quality improvement efforts. However, use of *behavioural tools* such as brainstorming, team-building, and nominal group decision-making, and *statistical methods* such as tables, control charts, and descriptive statistics, are also necessary for understanding and improving the processes.

The steps of statistical thinking necessary for increased understanding of and improvement in the processes are summarized in Fig. 1.3.

Figure 1.2
Quality Improvement Process Model

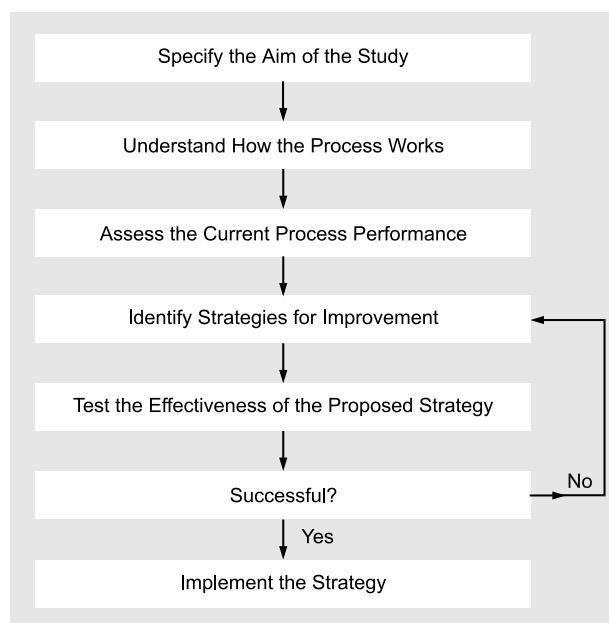
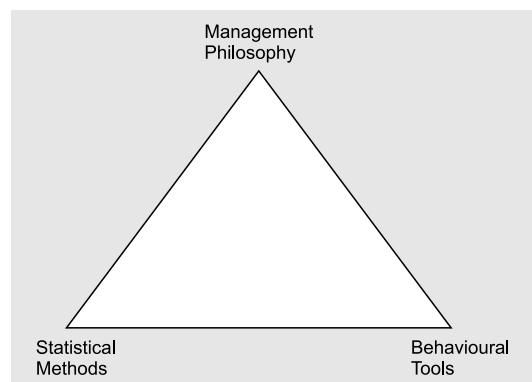


Figure 1.3
Flow Chart of Process Improvement

1.4 STATISTICS DEFINED

As Statistical Data The word statistics refers to a special discipline or a collection of procedures and principles useful as an aid in gathering and analysing numerical information for the purpose of drawing conclusions and making decisions. Since any numerical figure, or figures, cannot be called statistics owing to many considerations which decide its use, statistical data or mere data is a more appropriate expression to indicate numerical facts.

A few definitions which describe the characteristics of statistics are as follows:

- The classified facts respecting the condition of the people in a state . . . especially those facts which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement. —Webster

This definition is quite narrow as it confines the scope of statistics only to such facts and figures which are related to the conditions of the people in a state.

Quantitative data:

Numerical data measured on an interval or ratio scales to describe 'how much' or 'how many'.

- By statistics we mean **quantitative data** affected to a marked extent by multiplicity of causes. —Yule and Kendall
- By statistics we mean aggregates of facts affected to a marked extent by multiplicity of causes numerically expressed, enumerated, or estimated according to reasonable standards of accuracy, collected in a systematic manner for predetermined purpose and placed in relation to each other. —Horace Secrist

The definition given by Horace is more comprehensive than that of Yule and Kendall. This definition highlights the following important characteristics

- (i) statistics are aggregates of facts,
- (ii) statistics are effected to a marked extent by multiplicity of causes,
- (iii) statistics are numerically expressed,
- (iv) statistics are enumerated or estimated according to reasonable standards of accuracy,
- (v) statistics are collected in a systematic manner for a pre-determined purpose, and
- (vi) statistics are placed in relation to each other.

As Statistical Methods Methods adopted as aids in the collection and analysis of numerical information or statistical data for the purpose of drawing conclusions and making decisions are called *statistical methods*.

Statistical methods, also called statistical techniques, are sometimes loosely referred to cover 'statistics' as a subject in whole. There are two branches of statistics: (i) *Mathematical statistics* and (ii) *Applied statistics*. Mathematical statistics is a branch of mathematics and is theoretical. It deals with the basic theory about how a particular statistical method is developed. Applied statistics, on the other hand, uses statistical theory in formulating and solving problems in other subject areas such as economics, sociology, medicine, business/industry, education, and psychology.

The field of applied statistics is not easy because the rules necessary to solve a particular problem are not always obvious although the guiding principles that underlie the various methods are identical regardless of the field of their application. As a matter of fact, experience and judgment are otherwise more necessary to execute a given statistical investigation.

The purpose of this book is limited to discussing the fundamental principles and methods of applied statistics in a simple and lucid manner so that readers with no previous formal knowledge of mathematics could acquire the ability to use statistical methods for making managerial decisions.

A few relevant definitions of statistical methods are given below:

- Statistics is the science which deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry. —Seligman
- The science of statistics is the method of judging, collecting natural or social phenomenon from the results obtained from the analysis or enumeration or collection of estimates. —King

A. L. Bowley has given the following three definitions keeping in mind various aspects of statistics as a science:

- Statistics may be called the science of counting.
- Statistics may be called the science of average.
- Statistics is the science of the measurement of social organism regarded as a whole in all its manifestations.

These definitions confine the scope of statistical analysis only to 'counting, average, and application' in the field of sociology alone. Bowley realized this limitation and said that statistics cannot be confined to any science. Another definition of statistics given by Croxton and Cowden is as follows:

- Statistics may be defined as a science of collection, presentation, analysis and interpretation of numerical data.
- Croxton and Cowden

This definition has pointed out four stages of statistical investigation, to which one more stage ‘organization of data’ rightly deserves to be added. Accordingly, statistics may be defined as the science of collecting, organizing, presenting, analysing, and interpreting numerical data for making better decisions.

1.5 TYPES OF STATISTICAL METHODS

Statistical methods, broadly, fall into the following two categories:

- Descriptive statistics, and
- Inferential statistics

Descriptive statistics includes statistical methods involving the collection, presentation, and characterization of a set of data in order to describe the various features of that set of data.

In general, methods of descriptive statistics include graphic methods and numeric measures. Bar charts, line graphs, and pie charts comprise the graphic methods, whereas numeric measures include measures of central tendency, dispersion, skewness, and kurtosis.

Inferential statistics includes statistical methods which facilitate estimating the characteristics of a population or making decisions concerning a population on the basis of sample results. **Sample** and population are two relative terms. The larger group of units about which inferences are to be made is called the population or universe, while a sample is a fraction, subset, or portion of that universe.

Inferential statistics can be categorized as *parametric* or *non-parametric*. The use of parametric statistics is based on the assumption that the population from which the sample is drawn, is normally distributed. Parametric statistics can be used only when data are collected on an interval or ratio scale. Non-parametric statistics makes no explicit assumption regarding the normality of distribution in the population and is used when the data are collected on a nominal or ordinal scale.

The need for sampling arises because in many situations data are sought for a large group of elements such as individuals, companies, voters, households, products, customers, and so on to make inferences about the population that the sample represents. Thus, due to time, cost, and other considerations data are collected from only a small portion of the population called *sample*. The concepts derived from probability theory help to ascertain the likelihood that the analysis of the characteristics based on a sample do reflect the characteristics of the population from which the sample is drawn. This helps the decision-maker to draw conclusions about the characteristics of a large population under study.

Following definitions are necessary to understand the concept of inferential statistics:

- A *process* is a set of conditions that repeatedly come together to transform inputs into outcomes. Examples include a business process to serve customers, length of time to complete a banking transaction, manufacturing of goods, and so on.
- A *population* (or *universe*) is a group of elements or observations relating to a phenomenon under study for which greater knowledge and understanding is needed. The observations in population may relate to employees in a company, a large group of manufactured items, vital events like births and deaths or road accidents. A population can be *finite* or *infinite* according to the number of observations under statistical investigation.
- A *statistical variable* is an operationally defined characteristic of a population or process and represents the quantity to be observed or measured.
- A *sample* is a group of some, but not all, of the elements or observations of a population or process. The individual elements of a sample are called *sampling* or *experimental units*.
- A *parameter* is a descriptive or summary measure (a numerical quantity) associated with a statistical variable that describes a characteristic of the entire population.

Descriptive statistics: It consists of procedures used to summarize and describe the characteristics of a set of data.

Inferential statistics: It consists of procedures used to make inferences about population characteristics on the basis of sample results.

Sample: A subset (portion) of the population.

- A *statistic* is a numerical quantity that describes the characteristic of a sample drawn from a population.

For example, a manufacturer who produces electrical coils wants to learn the average resistance of coils. For this he selects a sample of coils at regular intervals of time and measures the resistance of each. If the sample average does not fall within the specified range of variations, the process controls are checked and adjustments are made. In this example, the population or universe would be all the coils being produced by the manufacturing process; the statistical variable is the resistance of a coil; statistic is the average resistance of coils in a given sample; parameters of interest are the average resistance and variation in resistance among manufactured coils; and sampling units are the coils selected for the sample.

1.6 IMPORTANCE AND SCOPE OF STATISTICS

The scope of application of statistics has assumed unprecedented dimensions these days. Statistical methods are applicable in diverse fields such as economics, trade, industry, commerce, agriculture, bio-sciences, physical sciences, education, astronomy, insurance, accountancy and auditing, sociology, psychology, meteorology, and so on. Bringing out its wide applications, Carroll D. Wright (1887), United States Commissioner of the Bureau of Labour, has explained the importance of statistics by saying:

To a very striking degree our culture has become a statistical culture. Even a person who may never have heard of an index number is affected by those index numbers which describe the cost of living. It is impossible to understand Psychology, Sociology, Economics or a Physical Science without some general idea of the meaning of an average, of variation, of concomitance of sampling, of how to interpret charts and tables.

In the recent past, statistics has acquired its importance as a subject of study in the curricula of many other disciplines. According to the statistician Bowley, '*A knowledge of statistics is like a knowledge of foreign language or of algebra, it may prove of use at any time under any circumstances*'.

Given below is a brief discussion on the importance of statistics in a few other important disciplines.

1.6.1 Statistics and the State

A state in the modern setup collects the largest amount of statistics for various purposes. It collects data relating to prices, production, consumption, income and expenditure, investments, and profits. Popular statistical methods such as time-series analysis, index numbers, forecasting, and demand analysis are extensively practised in formulating economic policies. Governments also collect data on population dynamics in order to initiate and implement various welfare policies and programmes.

In addition to statistical bureaus in all ministries and government departments in the Central and state governments, other important agencies in the field are the Central Statistical Organisation (CSO), National Sample Survey Organization (NSSO), and the Registrar General of India (RGI).

1.6.2 Statistics in Economics

Statistical methods are extensively used in all branches of economics. For example:

- (i) Time-series analysis is used for studying the behaviour of prices, production and consumption of commodities, money in circulation, and bank deposits and clearings.
- (ii) Index numbers are useful in economic planning as they indicate the changes over a specified period of time in (a) prices of commodities, (b) imports and exports, (c) industrial/agricultural production, (d) cost of living, and the like.
- (iii) Demand analysis is used to study the relationship between the price of a commodity and its output (supply).

- (iv) Forecasting techniques are used for curve fitting by the principle of least squares and exponential smoothing to predict inflation rate, unemployment rate, or manufacturing capacity utilization.

1.6.3 Statistics in Business Management

According to Wallis and Roberts, ‘Statistics may be regarded as a body of methods for making wise decisions in the face of uncertainty.’ Ya-Lin-Chou gave a modified definition over this, saying that ‘Statistics is a method of decision-making in the face of uncertainty on the basis of numerical data and calculated risks.’ These definitions reflect the applications of statistics in the development of general principles for dealing with uncertainty.

Statistical reports provide a summary of business activities which improves the capability of making more effective decisions regarding future activities. Discussed below are certain activities of a typical organization where statistics plays an important role in their efficient execution.

Marketing Before a product is launched, the market research team of an organization, through a pilot survey, makes use of various techniques of statistics to analyse data on population, purchasing power, habits of the consumers, competitors, pricing, and a host of other aspects. Such studies reveal the possible market potential for the product.

Analysis of sales volume in relation to the purchasing power and concentration of population is helpful in establishing sales territories, routing of salesman, and advertising strategies to improve sales.

Production Statistical methods are used to carry out R&D programmes for improvement in the quality of the existing products and setting quality control standards for new ones. Decisions about the quantity and time of either self-manufacturing or buying from outside are based on statistically analysed data.

Finance A statistical study through correlation analysis of profits and dividends helps to predict and decide probable dividends for future years. Statistics applied to analysis of data on assets and liabilities and income and expenditure helps to ascertain the financial results of various operations.

Financial forecasts, break-even analysis, investment decisions under uncertainty—all involve the application of relevant statistical methods for analysis.

Personnel In the process of manpower planning, a personnel department makes statistical studies of wage rates, incentive plans, cost of living, labour turnover rates, employment trends, accident rates, performance appraisal, and training and development programmes. Employer-employee relationships are studied by statistically analysing various factors—wages, grievances handling, welfare, delegation of authority, education and housing facilities, and training and development.

1.6.4 Statistics in Physical Sciences

Currently there is an increasing use of statistical methods in physical sciences such as astronomy, engineering, geology, meteorology, and certain branches of physics. Statistical methods such as sampling, estimation, and design of experiments are very effective in the analysis of quantitative expressions in all fields of most physical sciences.

1.6.5 Statistics in Social Sciences

The following definitions reflect the importance of statistics in social sciences.

- Statistics is the science of the measurement of social organism, regarded as a whole in all its manifestations. —Bowley
- The science of statistics is the method of judging collective, natural or social phenomenon from the results obtained from the analysis, enumeration or collection of estimates. —W. I. King

Some specific areas of applications of statistics in social sciences are as listed below:

- (i) Regression and correlation analysis techniques are used to study and isolate all those factors associated with each social phenomenon which bring out the changes in data with respect to time, place, and object.
- (ii) Sampling techniques and estimation theory are indispensable methods for conducting any social survey pertaining to any strata of society, and drawing valid inferences.
- (iii) In sociology, statistical methods are used to study mortality (death) rates, fertility (birth rates) trends, population growth, and other aspects of vital statistics.

1.6.6 Statistics in Medical Sciences

The knowledge of statistical techniques in all natural sciences—zoology, botany, meteorology, and medicine—is of great importance. For example, for proper diagnosis of a disease, the doctor needs and relies heavily on factual data relating to pulse rate, body temperature, blood pressure, heart beats, and body weight.

An important application of statistics lies in using the *test of significance* for testing the efficacy of a particular drug or injection meant to cure a specific disease. Comparative studies for effectiveness of a particular drug/injection manufactured by different companies can also be made by using statistical techniques such as the *t*-test and *F*-test.

To study plant life, a botanist has to rely on data about the effect of temperature, type of environment, and rainfall, and so on.

1.6.7 Statistics and Computers

Computers and information technology, in general, have had a fundamental effect on most business and service organizations. Over the last decade or so, however, the advent of the personal computer (PC) has revolutionized both the areas to which statistical techniques are applied. PC facilities such as spreadsheets or common statistical packages have now made such analysis readily available to any business decision-maker. Computers help in processing and maintaining past records of operations involving payroll calculations, inventory management, railway/airline reservations, and the like. Use of computer softwares, however, presupposes that the user is able to interpret the computer outputs that are generated.

Remark We discussed above the usefulness of statistical techniques in some important fields. However, the scope of statistics is not limited to these only. Statistical data and methods are useful to banking, research and development, insurance, astronomy, accountancy and auditing, social workers, labour unions, chambers of commerce, and so on.

1.7 LIMITATIONS OF STATISTICS

Although statistics has its applications in almost all sciences—social, physical, and natural—it has its own limitations as well, which restrict its scope and utility.

1.7.1 Statistics Does Not Study Qualitative Phenomena

Since statistics deals with numerical data, it cannot be applied in studying those problems which can be stated and expressed quantitatively. For example, a statement like 'Export volume of India has increased considerably during the last few years' cannot be analysed statistically. Also, qualitative characteristics such as honesty, poverty, welfare, beauty, or health, cannot directly be measured quantitatively. However, these subjective concepts can be related in an indirect manner to numerical data after assigning particular scores or quantitative standards. For example, attributes of intelligence in a class of students can be studied on the basis of their Intelligence Quotients (IQ) which is considered as a quantitative measure of the intelligence.

1.7.2 Statistics Does Not Study Individuals

According to Horace Secrist '*By statistics we mean aggregate of facts affected to a marked extent by multiplicity of factors . . . and placed in relation to each other.*' This statement implies that a single or isolated figure cannot be considered as statistics, unless it is part of the aggregate of facts relating to any particular field of enquiry. For example, price of a single commodity or increase or decrease in the share price of a particular company does not constitute statistics. However, the aggregate of figures representing prices, production, sales volume, and profits over a period of time or for different places do constitute statistics.

1.7.3 Statistics Can be Misused

Statistics are liable to be misused. For proper use of statistics one should have enough skill and experience to draw accurate and sensible conclusions. Further, valid results cannot be drawn from the use of statistics unless one has a proper understanding of the subject to which it is applied.

The greatest danger of statistics lies in its use by those who do not possess sufficient experience and ability to analyse and interpret statistical data and draw sensible conclusions. Bowley was right when he said that '*statistics only furnishes a tool though imperfect which is dangerous in the hands of those who do not know its use and deficiencies.*' For example, the conclusion that smoking causes lung cancer, since 90 per cent of people who smoke die before the age of 70 years, is statistically invalid because here nothing has been mentioned about the percentage of people who do not smoke and die before reaching the age of 70 years. According to W. I. King, '*statistics are like clay of which you can make a God or a Devil as you please.*' He also remarked, '*science of statistics is the useful servant but only of great value to those who understand its proper use.*'

1.8 HOW TO LIE WITH STATISTICS

Despite the happy use of statistics and statistical methods in almost every profession, it is still distrusted. Statistics is considered one of the three types of lies: lies, damn lies, and statistics. Listed below may be two reasons for such a notion being held by people about statistics.

- (i) Figures being innocent and convincing, are easily believable.
- (ii) Figures which support a particular statement may not be true. Such figures may be incomplete, inaccurate, or deliberately manipulated by prejudiced persons in an attempt to deceive the user or attain ones own motive.

Table 1.1 lists some of the personal qualities and attributes considered necessary for an individual to be an effective statistician:

Table 1-1 Personal Qualities and Attributes For A Statistician*

<i>An effective statistician</i>
<ul style="list-style-type: none"> • is well-trained in the theory and practice of statistics. • is an effective problem-solver. • has good oral and written communication skills. • can work within the constraints of real-life. • knows how to use computers to solve problems. • understands the realities of statistical practices. • has a pleasing personality and is able to work with others. • gets highly involved in solving organizational problems. • is able to extend and develop statistical methodology. • can adapt quickly to new problems and challenges. • produces high quality work in an orderly and timely fashion.

* Source: "Preparing Statistics for Cancers in Industry," *The American Statistician*, Vol. 34, No. 2, May 1980.

Conceptual Questions 1A

1. What is statistics? How do you think that the knowledge of statistics is essential in management decisions. Give examples.
2. Write a brief note on the application of statistics in business and industry.
3. Discuss the meaning and scope of statistics, bringing out its importance particularly in the field of trade and commerce.
4. (a) How far can statistics be applied for business decisions? Discuss briefly bringing out limitations, if any
 (b) Define 'statistics' and give its main limitations.
5. (a) Explain how statistics plays an important role in management planning and decision-making?
 (b) Define statistics and statistical methods. Explain the uses of statistical methods in modern business.
[Vikram Univ., MBA, 1996]
6. Statistical methods are the most dangerous tools in the hands of an inexpert. Examine this statement. How are statistics helpful in business and industry? Explain.
[Delhi Univ., MBA, 1999]
7. (a) Define statistics. Discuss its applications in the management of business enterprises. What are its limitation, if any.
[Jodhpur Univ., MBA; HP Univ., MBA, 1996]
 (b) Explain the utility of statistics as a managerial tool. Also discuss its limitations.
[Osmania Univ., MBA, 1998]
8. What role does Business Statistics play in the management of a business enterprise? Examine its scope and limitations.
[Delhi Univ., MBA, 1998]
9. (a) Statistics are like clay of which you can make a God or Devil, as you please. Explain.
 (b) There are three known lies : lies, dam-lies and statistics. Comment on this statement and point out the limitations of statistics.
10. Discuss briefly the applications of Business Statistics, pointing out their limitations, if any. [Delhi Univ., MBA, 1997]
11. Describe the main areas of business and industry where statistics are extensively used.
12. Statistics affects everybody and touches life at many points. It is both a science and an art. Explain this statement with suitable examples.
13. With the help of few examples explain the role of statistics as a managerial tool.
14. 'Statistics in the science of estimates and probabilities'. Explain the statement and discuss the role of statistics in the management of business enterprises.
15. Are statistical methods likely to be of any use to a business firm ? Illustrate your answer with some typical business problems and the statistical techniques to be used there.
[HP Univ., MBA, 1996; Delhi Univ., MBA, 2000;
Roorkee Univ., MBA, 2000]
16. 'Statistics is a body of methods for making wise decisions in the face of uncertainty'. Comment on the statement bringing out clearly how statistics helps in business decision-making.
[Osmania Univ., MBA, 1996]
17. 'Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write' Comment. Also give two examples, of how the science of statistics could be of use in managerial decision-making.
[HP Univ., MBA, 1996]
18. 'Statistics is a method of decision-making in the face of uncertainty on the basis of numerical data and calculated risks'. Comment and explain with suitable illustrations.
[Delhi Univ., MBA, 1992, 1993]
19. 'Without adequate understanding of statistics, the investigator in social sciences may frequently be like the blind man grouping in a dark closet for a black cat that is not there'. Comment. Give two examples of the use and abuse of statistics in business.
20. One can say that statistical inference includes an interest in statistical description as well, since the ultimate purpose of statistical inference is to describe population data. Does statistical inference differ from statistical description? Discuss.
21. What characteristics are inevitable in virtually all data and why is the understanding of it important?
22. 'Modern statistical tools and techniques are important for improving the quality of managerial decisions'. Explain this statement and discuss the role of statistics in the planning and control of business.
[HP Univ., MBA, 1998]
23. 'The fundamental gospel of statistics is to push back the domain of ignorance, rule of thumb, arbitrary or prepare decisions, traditions, and dogmatism, and to increase the domain in which decisions are made and principles are formulated on the basis of analysed quantitative facts'. Explain the statement with the help of a few business examples.
[Osmania Univ., MBA, 1999]
24. 'Statistics are numerical statements of facts but all facts numerically stated are not statistics'. Comment upon the statement.
25. (a) Define statistics. Why do some people look at this science with an eye of distrust?
 (b) 'The science of statistics is the most useful servant but only of great value to those who understand its proper use'. Discuss.
26. Bring out the applications of statistics in economics and business administration as a scientific tool. Also point out any two limitations of statistics.
[CA Foundation, May 1996]
27. Give an example of the use of descriptive statistics and inferential statistics in each of the following areas of application in a business firm.
 - (a) Production management
 - (b) Financial management
 - (c) Marketing management
 - (d) Personnel management
28. Discuss the differences between statistics as numerical facts and as a discipline or field of study.
29. ORG conducts weekly surveys of television viewing throughout the country. The ORG statistical ratings indicate the size of the viewing audience for each major network television programme. Rankings of the television

- programmes and of the viewing audience market shares for each network are published each week.
- (a) What is the organization, ORG, attempting to measure?

- (b) What is the population?
 (c) Why would a sample be used for this situation?
 (d) What kinds of decisions or actions are based on the ORG studies?

1.9 NEED FOR DATA

Statistical data are the basic material needed to make an effective decision in a particular situation. The main reasons for collecting data are as listed below:

- (i) To provide necessary inputs to a given phenomenon or situation under study.
- (ii) To measure performance in an ongoing process such as production, service, and so on.
- (iii) To enhance the quality of decision-making by enumerating alternative courses of action in a decision-making process, and selecting an appropriate one.
- (iv) To satisfy the desire to understand an unknown phenomenon.
- (v) To assist in guessing the causes and probable effects of certain characteristics in given situations.

For any statistical analysis to be useful, the collection and use of input data is extremely important. One can collect an enormous amount of data on a subject of interest in a compact and usable form from the internet. However, the reliability of such data is always doubtful. Thus, before relying on any interpreted data, either from a computer, internet or other source, we should study answers to the following questions: (i) Have data come from an unbaised source, that is, source should not have an interest in supplying the data that lead to a misleading conclusion, (ii) Do data represent the entire population under study i.e. how many observations should represent the population, (iii) Do the data support other evidences already available. Is any evidence missing that may cause to arrive at a different conclusion? and (iv) Do data support the logical conclusions drawn. Have we made conclusions which are not supported by data.

Nowadays computers are extensively used for processing data so as to draw logical conclusions. Since a computer is only a machine used for fast processing of input data, the output data received are only as accurate as the data that is fed in. The decision-maker thus needs to be careful that the data he is using comes from a valid source and evidences that might cause him to arrive at a different conclusion are not missing.

In order to design an experiment or conduct a survey one must understand the different types of data and their measurement levels.

1.9.1 Types of Data

Statistical data are the outcome of a continuous process of measuring, counting, and/or observing. These may pertain to several aspects of a phenomenon (or a problem) which are measurable, quantifiable, countable, or classifiable. While conducting a survey or making a study, an investigator develops a method to ask several questions to deal with the variety of characteristics of the given population or universe. These characteristics which one intends to investigate and analyse are termed as *variables*. The data, which are the observed outcomes of these variables, may vary from response to response. Consumer behaviour (attitude), profit/loss to a company, job satisfaction, drinking and/or smoking habits, leadership ability, class affiliation or status are examples of a variable.

Table 1.2 summarizes the types of variables which can be studied to yield the observed outcomes in relation to the nature of data, information, and measurement.

Table 1.2 Nature of Data, Information, and Measurement

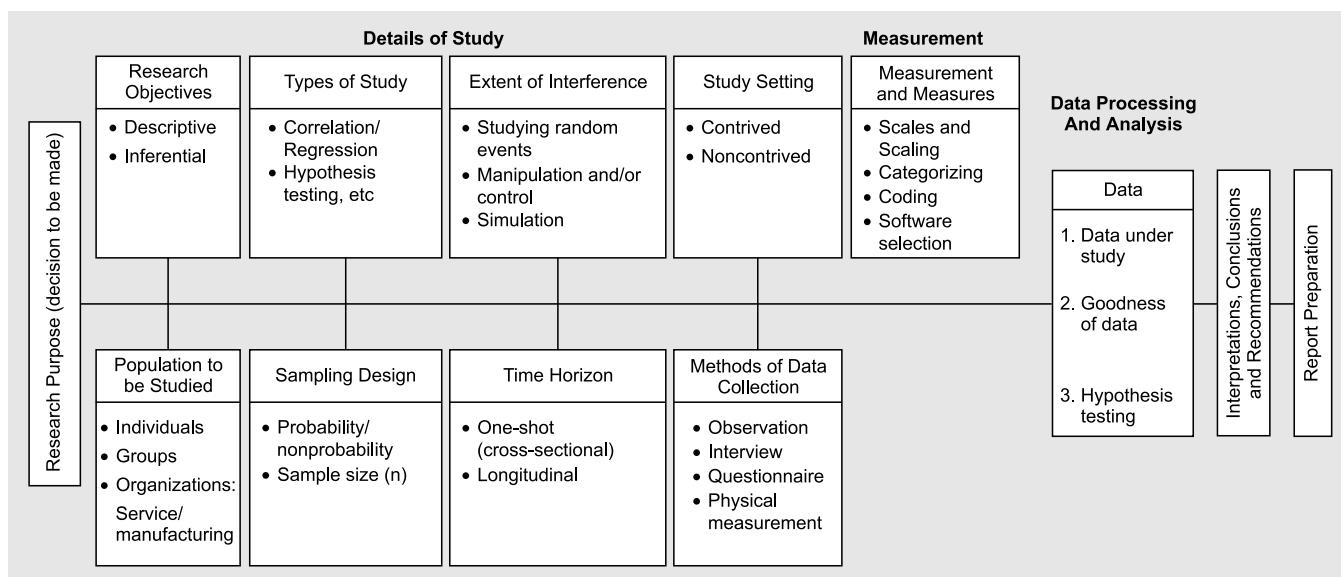
Data Type	Information Type	Measurement Type
Categorical	→ Do you practice Yoga?	Yes <input type="checkbox"/> No <input type="checkbox"/>
Numerical ↗ Discrete	→ How many books do you have in your library?	Number
	→ What is your height?	Centimetres or Inches

It may be noted from Table 1.2 that categorical variables are those which are not expressed in numerical terms. Sex, religion, and language are a few examples of such variables. The numerical variables are classified into two categories:

- Discrete variables—which can only take certain fixed integer values. The number of cars sold by Maruti Udyog Ltd. in 2001, or the number of employees in an organization are examples of discrete variables.
- Continuous variables—which can take any numerical value. Measurement of height, weight, length, in centimetres/inches, grams/kilograms are a few examples of continuous variables.

Remark: *Discrete data* are numerical measurements that arise from a process of counting, while *continuous data* are numerical measurements that arise from a process of measuring.

A flow chart of the research process is shown in Fig. 1.4.



1.10 PRINCIPLES OF MEASUREMENT

Just as there are rules or guidelines that have to be followed to ensure that the wording of the questionnaire is appropriate to minimize bias, so also are some principles of measurement that are to be followed to ensure that the data collected are appropriate to test our hypothesis. These principles of measurement encompass the scales and scaling techniques used in measuring concepts, as well as the assessment of reliability and validity of the measures used. Appropriate scales have to be used depending on the type of data that need to be obtained. Once data are obtained, the “goodness of data” is assessed through tests of validity and reliability. Validity establishes how well a technique, or a process measures a particular concept, the reliability indicates how stably and consistently the technique measures the variable.

In general, the principles of measurement (scaling) has three characteristics:

- Numbers are ordered. One number is less than, equal to or greater than another number.
- Difference between numbers are ordered. The difference between any pair of numbers is greater than, less than or equal to the difference between any other pair of numbers.
- The number series has a unique origin indicated by the number zero.

The combinations of these characteristics of *order*, *distance* and *origin* provide the following widely used classification of measurement scales:

<i>Scale of Measurement</i>	<i>Characteristics of Measurement</i>	<i>Basic Empirical Operation</i>
• Nominal	No order, distance or unique origin	Determination of categorical information. Numbers only identify groups which cannot be ordered
• Ordinal	Order but no distance or unique origin	Determination of greater or lesser values. Numbers allow ranking but no arithmetic
• Interval	Both order and distance but not unique	Determination of equality of intervals or differences. Intervals between numbers are meaningful
• Ratio	Order, distance and unique origin	Determination of equality of ratios. Intervals between numbers are meaningful and also their ratio as the lowest value is a meaningful zero.

Nominal Scale In nominal scaling the numerical values are either named or categorized in such a way that these values are mutually exclusive and collectively exhaustive. For example, shirt numbers in a football or cricket match are measured at a nominal level. A player wearing a shirt number 24 is not more of anything than a player wearing a shirt number 12 and is certainly not twice the number 12. In other words, if we use numbers to identify categories, they are recognised as levels only and have no qualitative value.

Nominal classification consists of any other number to separate groups if such groups are mutually exclusive and collectively exhaustive. For example, based on a nominal scale: each of the respondent has to fit into one of the six categories of nationality and scale will allow computation of the percentage of respondents who fit into each of these six categories

- Indian
- Nepalise
- Pakistani
- Srilankan
- Bhootanis
- Others

Nominal scale is said to be least powerful among four scales because this scale suggest no order or distance relationship and have no arithmetic origin. Few examples of nominal scaling are: sex, blood type, religion, nationality, etc.

Nominal scale is usually used for obtaining personal data such as gender, place of work, and so on, where grouping of individuals or objects is useful, as illustrated below.

- | | |
|----------------|-----------------------|
| 1. Your gender | 2. Your place of work |
| • Male | • Production |
| • Female | • Sales |
| | • Finance |
| | • Personnel |

Ordinal Scale In ordinal scaling the numerical values are categorised to denote qualitative differences among the various categories as well as rank-ordered in some meaningful way according to some preference. The preferences would be ranked from best to worst, first to last, numbered 1, 2, and so on.

The ordinal scale not only indicates the differences in the given items but also gives some information as to how respondents distinguish among these items by rank ordering them. However, the ordinal scale does not give any indication of the magnitude of the differences among the ranks, i.e. this scale implies a statement of 'greater than' or 'less than' (an equality statement is also acceptable) without stating how much greater or less. The real difference between ranks 1 and 2 may be more or less than the difference between ranks 4 and 5. The interval between values is not interpretable in an ordinal measure.

Nominal scale: A scale of measurement for a variable that uses a label (or name) to identify an attribute of an element of the data set.

Ordinal scale: A scale of measurement for a variable that is used to rank (or order) observations in the data set.

Besides 'greater than' and 'less than' measurements, other measurements such as 'superior to', 'happier than' or 'above' may also be used as ordinal scale.

Ordinal scale is usually used to rate the preference or usage of various brands of a product by individuals and to rank individuals, objects, or events. For example, rank the following personal computers with respect to their usage in your office, assigning the number 1 to the most used system, 2 to the next most used, and so on. If particular system is not used at all in your office, put a 0 against it.

IBM/AT	Compaq
IBM/XT	AT&T
Apple II	Tandy 2000
Macintosh	Zenith

Interval scale: A scale of measurement for a variable in which the interval between observations is expressed in terms of a fixed standard unit of measurement.

Interval Scale An interval scale allows us to perform certain arithmetical operations on the data collected from the respondents. Whereas the nominal scale only allows us to qualitatively distinguish groups by categorizing them into mutually exclusive and collectively exhaustive sets, the ordinal scale allows us to rank-order the preferences, and the interval scale allows us to compute the mean and the standard deviation of the responses on the variables. In other words, the interval scale not only classifies individuals according to certain categories and determines order of these categories; it also measures the magnitude of the differences in the preferences among the individuals.

In interval measurement the distance between attributes does have meaning. For example, when we measure temperature (in Fahrenheit), the distance from 30–40 is same as distance from 70–80. The interval values are interpretable. Because of this, it makes sense to compute an average of an interval variable, where it doesn't make sense to do so for ordinal scales.

Interval scale is used when responses to various questions that measure a variable can be determined on a five-point (or seven-point or any other number of points) scale. For example, respondents may be asked to indicate their responses to each of the questions by circling the number that best describes their feeling.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
	1	2	3	4	5
1. My job offers me a chance to test my abilities.	1	2	3	4	5
2. Mastering this job meant a lot to me.	1	2	3	4	5
3. Doing this job well is a reward in itself.	1	2	3	4	5
4. Considering the time spent on the job, I feel thoroughly familiar with my tasks and responsibilities.	1	2	3	4	5

Ratio scale: A scale of measurement for a variable that has interval which measurable in standard unit of measurement and a meaningful zero, i.e. the ratio of two values is meaningful.

Ratio Scale The ratio scale has an absolute measurement point. Thus the ratio scale not only measures the magnitude of the differences between points on the scale but also provides the proportions in the differences. It is the most powerful of the four scales because it has a unique zero origin. For example, a person weighing 90 kg is twice as heavy as one who weighs 45 kg. Since multiplying or dividing both of these numbers (90 and 45) by any given number will preserve the ratio of 2 : 1, the measure of central tendency of the ratio scale could either be arithmetic or geometric mean and the measure of dispersion could either be standard deviation or variance, or coefficient of variation.

Ratio scales are usually used in organizational research when exact figures on objective (as opposed to subjective) factors are desired. Few examples are as under:

1. How many other organizations did you work for before joining this job?
2. Please indicate the number of children you have in each of the following categories:

- over 6 years but under 12
- 12 years and over

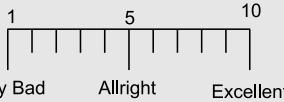
3. How many retail outlets do you operate?

The responses could range from 0 to any figure.

Graphic Rating Scale A graphical representation helps the respondent to indicate the response to a particular question by placing a mark at the appropriate point on the line as in the adjoining example.

Itemized Rating Scale This scale helps the respondent to choose one option that is most relevant for answering certain questions as in the following examples.

On a scale of 0 to 10 how would you rate your supervisor?



(a)		Not at all interested	Somewhat interested	Moderately interested	Very much interested
	How would you rate your interest in changing organizational policies?	1	2	3	4
(b)		Extremely Poor	Rather Poor	Quite Well	Very Well
	How well is the new distribution channel working?	1	2	3	4
					5

Other Measurement Scales

(a) Continuous rating scales

Type A

0	10	20	30	40	50	60	70	80	90	100
Unfavourable										Favourable

Type B

Unfavourable										Favourable
--------------	--	--	--	--	--	--	--	--	--	------------

(b) Itemized rating scale

Type A

Favourable										Unfavourable
_____ : _____ : _____ : _____ : _____ : _____ : _____										
extremely	quite	slightly	neither	slightly	quite	extremely				

Type B

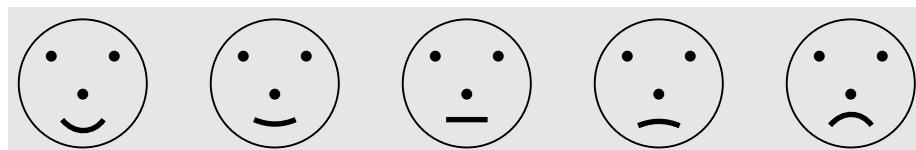
Favourable										Unfavourable
_____ : _____ : _____ : _____ : _____ : _____ : _____										

Type C

Favourable										Unfavourable
_____ : _____ : _____ : _____ : _____ : _____ : _____										

Type D

Favourable										Unfavourable
7	6	5	4	3	2	1				



(c) **Stapel scale**

Perfectly 7 6 5 4 3 2 1 Not at all

For example describe your visit to Shimla during January

safe	_____	boring	_____
pleasant	_____	status	_____
risky	_____	enjoyable	_____
necessary	_____	old	_____
useless	_____	valuable	_____
attractive	_____	cold	_____

Similarly, given on the next page are five characteristics of an automobile. Allocate 100 points among the characteristics such that the allocation represents the importance of each characteristic to you. The more points a characteristic receives, the more important it is. If the characteristic is not at all important, it is possible to assign zero points. If a characteristic is twice as important as some other, then it should receive twice as many points.

Characteristics	Number of Points
• Styling	50
• Ride	10
• Petrol mileage	35
• Warranty	5
• Closeness to dealer	<u>0</u>
	100

(d) **Semantic differential scale**

Describe going to Delhi during the summer vacations:

important	_____	_____	_____	_____	_____	_____	_____	unimportant
worthless	_____	_____	_____	_____	_____	_____	_____	valuable
good	_____	_____	_____	_____	_____	_____	_____	bad
rewarding	_____	_____	_____	_____	_____	_____	_____	punishing
useful	_____	_____	_____	_____	_____	_____	_____	useless
pessimistic	_____	_____	_____	_____	_____	_____	_____	optimistic
hard	_____	_____	_____	_____	_____	_____	_____	soft
boring	_____	_____	_____	_____	_____	_____	_____	interesting
active	_____	_____	_____	_____	_____	_____	_____	passive
compulsory	_____	_____	_____	_____	_____	_____	_____	voluntary
serious	_____	_____	_____	_____	_____	_____	_____	humorous
pleasant	_____	_____	_____	_____	_____	_____	_____	unpleasant

1.11 SOURCES OF DATA

The choice of a data collection method from a particular source depends on the facilities available, the extent of accuracy required in analyses, the expertise of the investigator, the time span of the study, and the amount of money and other resources required for data collection. When the data to be collected are very voluminous and require huge amounts of money, manpower, and time, reasonably accurate conclusions can be drawn by observing even a small part of the population provided the concept of sampling is used objectively.

Data sources are classified as (i) primary sources, and (ii) secondary sources.

1.11.1 Primary Data Sources

Individuals, focus groups, and/or panels of respondents specifically decided upon and

set up by the investigator for data collection are examples of primary data sources. Any one or a combination of the following methods can be chosen to collect primary data:

- (i) Direct personal observations
- (ii) Direct or indirect oral interviews
- (iii) Administrating questionnaires

The methods which may be used for primary data collection are briefly discussed below:

Observation In observational studies, the investigator does not ask questions to seek clarifications on certain issues. Instead, he records the behaviour, as it occurs, of an event in which he is interested. Sometimes mechanical devices are also used to record the desired data.

Studies based on observations are best suited for researches requiring non-self report descriptive data. That is, when respondents' behaviours are to be understood without asking them to part with the needed information. Diverse opinions in the diagnosis of a particular disease could be an example of an observational study.

Certain difficulties do arise during the collection of such data on account of (i) the observer's training, philosophy, opinions, and expectations, (ii) the interdependence of observations and inferences, and (iii) the inadequacies of the sense organs causing significant variations in the observations of the same phenomenon.

Interviewing Interviews can be conducted either face-to-face or over telephone. Such interviews provide an opportunity to establish a rapport with the interviewer and help extract valuable information. Direct interviews are expensive and time-consuming if a big sample of respondents is to be personally interviewed. Interviewers' biases also come in the way. Such interviews should be conducted at the exploratory stages of research to handle concepts and situational factors.

Telephonic interviews help establish contact with interviewers spread over distantly separated geographic locations and obtain responses quickly. This method is effective only when the interviewer has specific questions to ask and the needs responses promptly. Since the interviewer in this case cannot observe the non-verbal responses at the other end, the respondent can unilaterally terminate the interview without warning or explanation.

Questionnaire It is a formalized set of questions for extracting information from the target respondents. The form of the questions should correspond to the form of the required information. The three general forms of questions are: *dichotomous* (yes/no response type), *multiple choice*, and *open-ended*. A questionnaire can be administered personally or mailed to the respondents. It is an efficient method of collecting primary data when the investigator knows what exactly is required and how to measure such variables of interest as:

- Behaviour—past, present, or intended.
- Demographic characteristics—age, sex, income, and occupation.
- Level of knowledge.
- Attitudes and opinions.

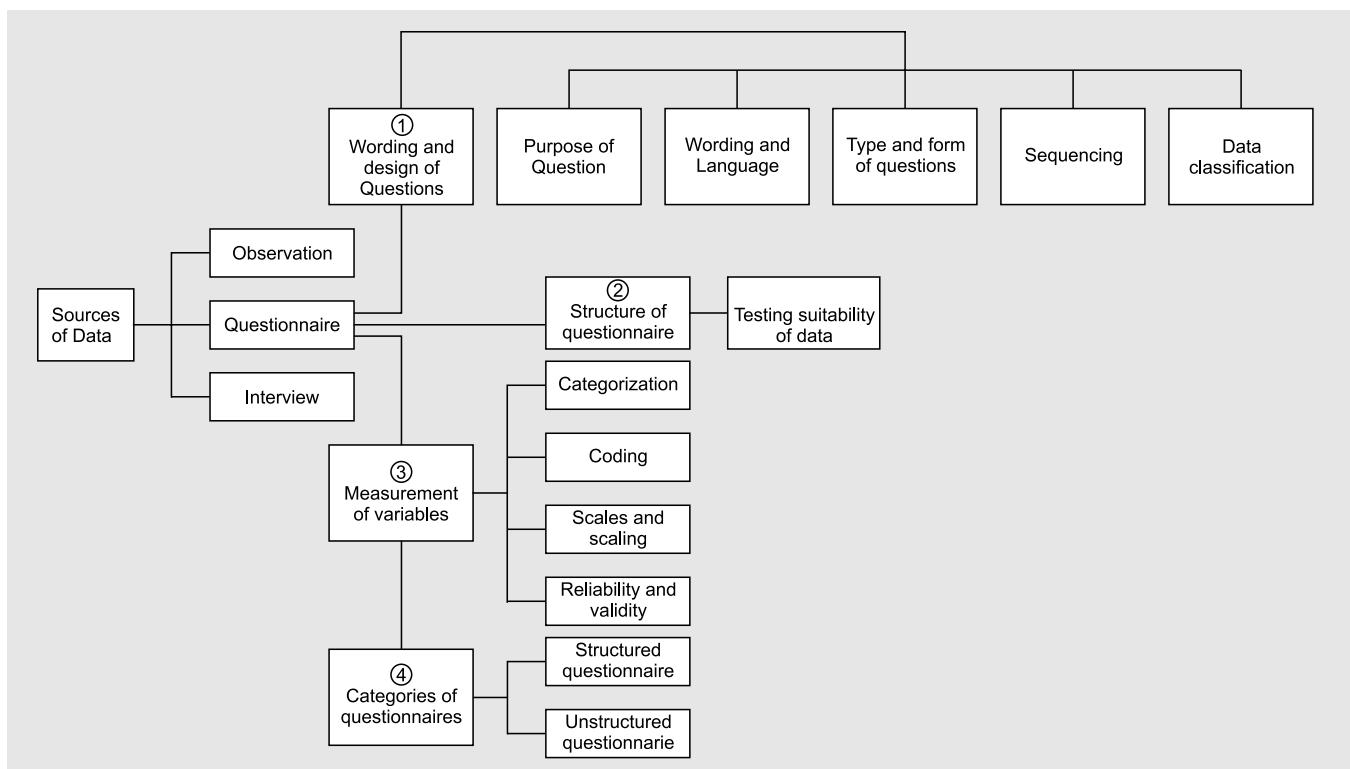
As such there are no set principles that must be used to design a questionnaire. However, general principles of questionnaire design based on numerous studies and experiences of survey researchers are shown in Fig. 1.5. A good questionnaire does, however, require the application of common sense, concern for the respondent, a clear concept of the information needed, and a thorough pre-testing of the questionnaire.

Questionnaire: A set of questions for extracting information from the target respondents.

1. The wording and design of questions The writing of good questions is an art, and a time-consuming art at that! In order to obtain valid and reliable responses one needs well-worded questions. There are a number of pitfalls to be avoided:

- *Open Ended Versus Closed Questions:* Open-ended questions allow respondents to answer them in any way they choose. Examples of open-ended questions are :
 - (i) State five things that are interesting and challenging in the job,
 - (ii) What you like about your supervisors or work environment,
 - (iii) What is your opinion about investment portfolio of your organization.

Figure 1.5
Principles of Questionnaire Design



A *Closed* question, would ask the respondents to make choices among a set of alternatives. For instance, instead of asking the respondent to state any five aspects of the job that are interesting and challenging, the researcher might list ten or fifteen characteristics that might seem interesting or challenging in jobs and ask the respondent to rank the first five among these.

Closed questions help the respondent to make quick decision by making a choice among the several alternatives that are provided. They also help the researcher to code the information easily for subsequent analysis. Of course, care has to be taken to ensure that the alternatives are mutually exclusive and collectively exhaustive. If there are overlapping categories, or if all possible alternatives are not given (i.e., the categories are not exhaustive), the respondents might get confused and the advantage of making a quick decision may be lost.

- *Positively and Negatively Worded Questions:* Instead of phrasing all questions positively, it is advisable to include some negatively worded questions also, so that it minimizes the tendency in respondents to mechanically circle the points toward one end of the scale. For example, a set of six questions are used to measure the variable ‘perceived success’ on a five-point scale, with 1 being ‘very low’ and 5 being ‘very high’ on the scale. A respondent who is not particularly interested in completing the questionnaire is more likely to stay involved and remain alert while answering the questions when positively and negatively worded questions are interspersed in the questionnaire. For instance, if the respondent had circled 5 for a positively worded question such as, ‘I feel I have been able to accomplish a number of different things in my job’, he cannot circle number 5 again to the negatively worded questions, ‘I do not feel I am very effective in my job.’ The use of double negatives excessively tends to confuse respondents. For instance, it is better to say ‘Coming to work is not great fun’ than to say ‘Not coming to work is greater fun than coming to work.’ Likewise, it is better to say ‘The strong people need no tonics’ than to say ‘Only the strong should be given no tonics.’
- *Double-Barreled Questions:* A question that lends different possible answers to its sub-parts is called a double-barreled question. Such questions should be avoided and

two or more separate questions should be asked. For example, the question “Do you think there is a good market for the product and that it will sell well?” could bring a ‘yes’ response to the first part (i.e., there is a good market for the product) and a ‘no’ response to the latter part (i.e., it will not sell well—for various other reasons). In this case, it would be better to ask two questions such as: (a) ‘Do you think there is a good market for the product?’ (b) ‘Do you think the product will sell well?’

- *Ambiguous Questions:* Questions that can be interpreted differently by different respondents should be avoided. For example, for the question such as: ‘To what extent would you say you are happy?’, the respondent might not be sure whether the question refers to his feelings at the workplace, or at home, or in general. Because, respondent might presume that the question relates to the workplace. Yet the researcher might have intended to inquire about the overall degree of satisfaction that the respondent experiences in everyday life—a feeling not specific to the workplace alone or at home.
- *Level of Wording:* It is important to tailor the level of wording of questions in accordance with the understanding of respondents. Jargons are to be avoided, and it should be established in the pilot study that the respondents understand the concepts. For instance, asking questions about ‘Trisomy 21’ might be inappropriate while ‘mongolism’ or ‘Down syndrome’ could be an intelligible. Use of double negatives should be avoided. In general, the questions should be simple and concise.
- *Biased and Leading Questions:* The wording of the questions should not lead the respondent to feel committed to respond in a certain way. For example, the question ‘How often do you go to church?’ may lead the respondent to respond in a way that is not entirely truthful if they, in fact, never go to church. Not only can the wording of a question be leading but the response format may also be leading. For example, if a ‘never’ response were excluded from the available answers to the above question, the respondent would be led to respond in an inaccurate way.

Bias might also arise from possible carry-over effects from answering a pattern of questions. For instance, a questionnaire on health workers’ attitudes to abortion might include the questions ‘Do you value human life?’ followed by ‘Do you think unborn babies should be murdered in their mothers’ wombs?’. In this case, the respondent is being led both by the context in which the second question is asked and the bias involved in the emotional wording of the questions. Surely, one would have to be a monster to answer ‘yes’ to the second question, given the way it was asked.

Finally, it should be kept in mind that even a good questionnaire might be invalidly administered. For instance, a survey on ‘Attitudes to migration’ might be answered less than honestly by respondents if the interviewer is obviously of immigrant background.

2. The structure of questionnaire A questionnaire may be structured in different ways, but typically the following components are included:

- *Introductory Statement:* The introductory statement describes the purpose of the questionnaire, the information sought, and how it is to be used. It also introduces the researchers and explains whether the information is confidential and/or anonymous.
- *Demographic Questions:* It is usual to collect information about the respondents, including details such as age, sex, education, and so on. It is best to position these questions first as they are easily answered and serve as a ‘warm-up’ to what follows.
- *Factual Questions:* It is generally easier for respondents to answer direct factual questions such as, ‘Do you have a driver’s licence?’ than to answer opinion questions. Often, this type of question is positioned early on in the questionnaire—also to help ‘warm up’!
- (iv) *Opinion Questions:* Questions that require reflection on the part of the respondent are usually positioned after the demographic and factual questions.
- *Closing Statements and Return Instructions:* The closing statements in a questionnaire usually thank the respondent for their participation, invite the respondent to take up any issues they feel have not been satisfactorily addressed in the questionnaire, and provide information on how to return the questionnaire.

It is best to avoid complicated structures involving, for example, many conditional questions such as 'If you answered yes to Question 6 and no to Question 9, please answer Question 10'. Conditional questions usually confuse respondents and ought to be avoided wherever possible.

3. Categories of questionnaires

- *Structured Questionnaire:* It is a formal list of questions to be posed to the respondents in a predetermined order. The responses permitted are also completely predetermined. Such questions are often called *closed questions* since the respondents are asked to make choices among a set of alternatives given by the investigator.

A structured questionnaire can also be *disguised* and *non-disguised*. This classification is based on whether the objectives of the study are disclosed or not disclosed to the respondents. A *structured undisguised questionnaire* is one where the purpose of the study and the particulars of the sponsor are disclosed to the respondent. In such cases, the questionnaire contains a list of questions in a predetermined order and freedom of response is limited only to the stated alternatives. Such questions help the respondent to make quick decisions by making a choice among the given alternatives. The alternatives provided have to be mutually exclusive and collectively exhaustive.

In the case of a *structured disguised questionnaire*, the objectives of the study and its sponsor are not disclosed to the respondents. Such questionnaires are not often used because it is felt necessary to have the respondents taken into confidence so that they appreciate the relevance of the desired information needed and willingly offer accurate answers.

- *Unstructured Questionnaire:* In this case, the investigator does not offer a limited set of response choices, but provides only a frame of reference within which the respondents are expected to answer. Such questionnaires are sometimes referred to as *open-ended questions*. Examples of open-ended questions are:

- (i) State three things that are interesting and challenging in your job.
- (ii) Write about the behaviour of a supervisor or the work environment.

These questions encourage the respondents to share as much information as possible in a free environment. The investigator may also provide extra guidance to the respondents by using a set of questions to promote discussion and elaboration.

The unstructured questionnaire is used in exploratory research studies or where the investigator is dealing with a complex phenomenon which does not lend itself to structured questioning. Such questionnaires are also useful when the investigator requires to know the respondent's emotions, needs, motivation level, attitude, and values. Obviously, using a questionnaire of this type, needs more time per interview and, therefore, raises the cost of the study. Editing and tabulation of these questionnaires also impose practical difficulties. Interestingly, unstructured questionnaires could also be of two types—*disguised* and *undisguised*.

Examples of questionnaire design

Two sets of questionnaire having most of the qualities of a good questionnaire are as under:

Questionnaire 1: Consumer Preferences

Name: _____ Age: _____

Address: _____

City: _____ Pin: _____ Phone: _____

Marital status: Married Single Occupation: _____

Family type: Joint Nuclear

Family members: Adults Children

Family income: Less than 10,000 10,000 to 15,000

15,000 to 20,000 More than 20,000

Remarks (if any):

Place and Date:

1. What kind of food do you normally eat at home?
 North Indian South Indian Mughlai Chinese
 Continental Italian Fast Food Others _____
2. How frequently do you eat out?
 In a week Once Twice Thrice More than thrice
 In a fortnight Once Twice Thrice More than thrice
 In a month Once Twice Thrice More than thrice
3. You usually go out with:
 Family Friends Colleagues Others _____
4. Any specific days when you go out:
 Weekdays Weekends Holidays Special Occasions
 No specific days
5. You generally go out for:
 Lunch Snacks Dinner Party/Picnics
 Others _____
6. Where do you usually go:
 Restaurant Chinese joint Fast food joint Others _____
7. Who decides on the place to go:
 Husband Wife Children Others _____
8. How much do you spend on eating out (one time):
 Below 200 200–500 500–800 More than 800
9. What are the factors that determine your choice for the restaurant/joint?
 Rank the following from 1–9 (9-highest score):
 Restaurant Chinese joint Fast food joint Others _____
10. Name the fast food giants that you are aware of (in Delhi):
 Nirula's Wimpy's McDonalds Pizza Hut
 Domino's Slice of Italy Pizza Express Others _____
11. How frequently do you go out/order for fast food?
 Very frequently Often Sometimes Never
12. What do you prefer: Going Out Home Delivery Take Away
13. Which of the places mentioned above in Q.10 are visited by you (and why):
 (a) Most frequently _____
 (b) Sometimes _____
 (c) Never _____
14. What are the distinguishing factors you look for in fast food service:
 (Rank from 1 to 8, 8-highest score)
 Quality Service Location Wide Menu Range
 Price Taste Home Delivery Others _____
15. What your order normally consists of:
 Pizza Burgers Footlong Curries & Breads
 Soups Pasta Desert Others _____
16. The price paid by you for the above is:

<i>Outlets</i>	<i>Very High</i>	<i>Little High</i>	<i>Just Right</i>
Nirula's			
Wimpy's			
Pizza Hut			
Domino's			
Slice of Italy			
Pizza Express			
McDonalds			
Others			

17. If you feel that the price paid by you is very high, what should be the price according to you:

Items	Vegetarian	Non-Vegetarian
Pizza		
Burger		
Footlong		
Others		

Questionnaire 2: Journal outlets for Production/Operations Management (POM) Research

If you have *not received* a Ph.D. degree, and have *not accepted* a full-time teaching position yet, mark the tick (✓) and stop. You need not complete the questionnaire.

If you have *not received* a Ph.D. degree but have *accepted* a full-time teaching position somewhere, mark the tick (✓). Skip Question 11, and answer all other questions.

1. How relevant do you consider the journal as a Production/Operations Management (POM)-related research outlet? Use the scale below.

1	2	3	4	5	6	7	8	9
Most relevant	Quite relevant	Relevant			Somewhat relevant			Not relevant

2. Based on the quality of the POM-related articles published, how would you rate the journal? (Use the scale below)

1	2	3	4	5	6	7	8	9	0
Level	Level	Level	Level	Level	Level	Level	Level	Level	Not possible to rate
A	A ⁻	B	B ⁻	C					

3. How does your institution/college rate this journal? (Use the scale in Question 2)
 4. How many articles have you authored or co-authored in this journal (include any articles that are currently in the press)?

Academy of Management Journal _____

Academy of Management Review _____

Computers and Industrial Engineering _____

Computers and Operations Research _____

Decision Sciences _____

European Journal of Operational Research _____

Harvard Business Review _____

Interfaces _____

Journal of Operations Management _____

Journal of Operational Research Society _____

Journal of Purchasing and Materials Management _____

Management Science _____

Naval Research Logistics Quarterly _____

Omega _____

Operations Research _____

Production and Inventory Management _____

Production and Operations Management _____

(List below any other journal that you consider related to POM research)

5. Using the scale below, please indicate the importance of the following factors in your assessment of the quality of a POM journal.

1	2	3	4	5	6	7	8	9
Most important	Very important	Important			Somewhat important	Not important at all		

____ Acceptance rate	____ Number of issues per year
____ Methodological rigour of the published work	____ Age of the journal
____ Editor and editorial board members	____ Professional organization that sponsors the journal
____ Authors who publish in the journal	____ Others (please specify)

6. At this stage of your career, how important for your career advancement is the quality of the journals in which your articles appear? (Use the scale below)

1	2	3	4	5	6	7	8	9
Most important	Very important	Important			Somewhat important	Not important at all		

7. At this stage of your career, how important for your career advancement is the quality of articles you author/co-author? (use the scale below)

1	2	3	4	5	6	7	8	9
Most important	Very important	Important			Somewhat important	Not important at all		

8. How much weightage does your institution/college place on research and publication in evaluating your annual performance? _____ (use a number between 0 and 100%)

9. What business degree(s) is (are) offered by the institution in which you teach? (tick all that apply)

- Undergraduate Masters level (MBA, MCA, M.Tech, etc.)
 Doctoral (M.Phil, Ph.D.)

10. What is your academic rank?

- Full professor Associate professor Assistant professor
 Other (e.g., instructor, lecturer, etc.)

11. In which year was your Ph.D. degree granted? _____

12. How many POM-related articles have you authored/co-authored in referenced journals? (include any articles that are currently in the press) _____

1.11.2 Secondary Data Sources

Secondary data refer to those data which have been collected earlier for some purpose other than the analysis currently being undertaken. Besides newspapers and business magazines, other sources of such data are as follows:

1. External secondary data sources

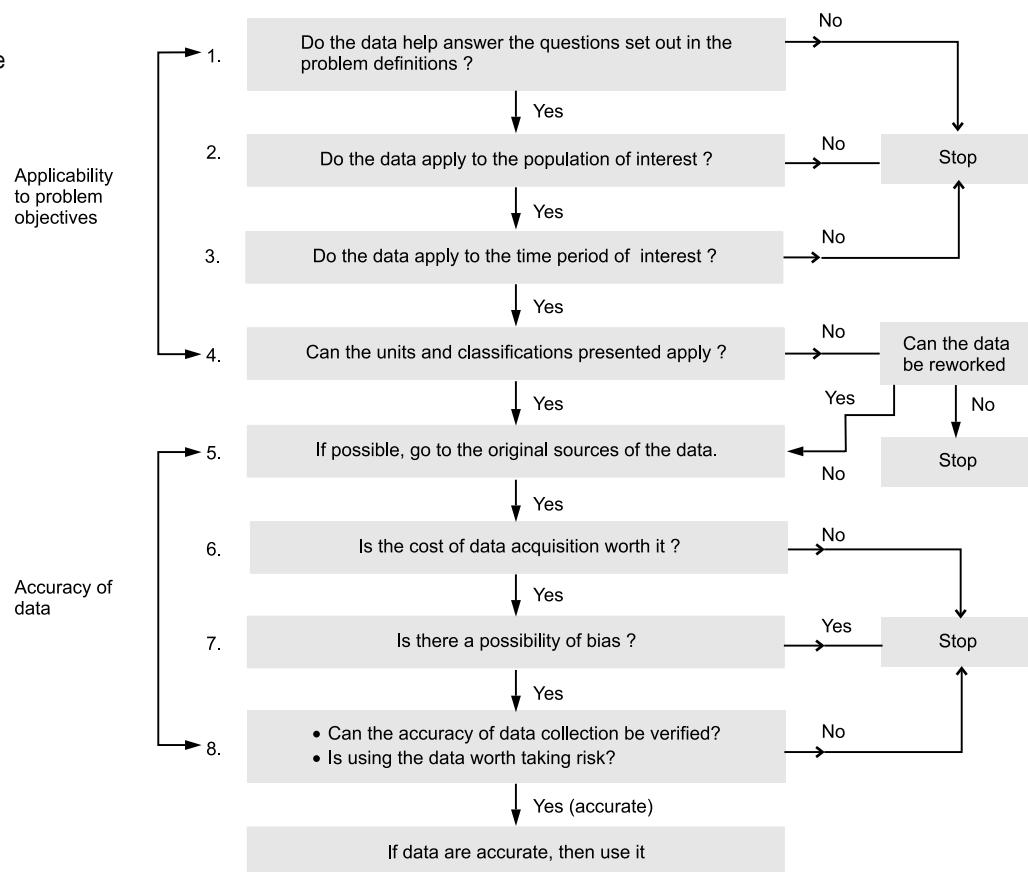
- Government publications, which include
 - (i) The National Accounts Statistics, published by the Central Statistical Organization (CSO). It contains estimates of national income for several years, growth rate, and rate on major economic activities such as agriculture, industry, trade, transport, and so on;
 - (ii) Wholesale Price Index, published by the office of the Economic Advisor, Ministry of Commerce and Industry;
 - (iii) Consumer Price Index;
 - (iv) Reserve Bank of India bulletins;
 - (v) Economic Survey.
- Non-Government publications include publications of various industrial and trade associations such as
 - (i) The Indian Cotton Mills Association

- (ii) The various Chambers of Commerce
- (iii) The Bombay Stock Exchange, which publishes a directory containing financial accounts, key profitability and other relevant data.
- Various syndicate services such as Operations Research Group (ORG). The Indian Market Research Bureau (IMRB) also collects and tabulates abundant marketing information to suit the requirements of individual firms, making the same available at regular intervals.
- International organizations which publish data are:
 - (i) The International Labour Organization (ILO)—which publishes data on the total and active population, employment, unemployment, wages, and consumer prices.
 - (ii) The Organization for Economic Cooperation and Development (OECD)—which publishes data on foreign trade, industry, food, transport, and science and technology.
 - (iii) The International Monetary Fund (IMF)—which publishes reports on national and international foreign exchange regulations and other trade barriers, foreign trade, and economic developments.

2. Internal secondary data sources The data generated within an organization in the process of routine business activities, are referred to as internal secondary data. Financial accounts, production, quality control, and sales records are examples of such data. However, data originating from one department of an organization may not be useful for another department in its original form. It is, therefore, desirable to condense such data into a form needed by the other.

Figure 1.6

Flow Chart Showing the Procedure for Evaluating Secondary Data



Advantages and Disadvantages of Secondary Data

Secondary data have their own advantages and disadvantages. The advantages are that such data are easy to collect and involve relatively lesser time and cost. Deficiencies and gaps can be identified easily and steps can be taken promptly to overcome the same.

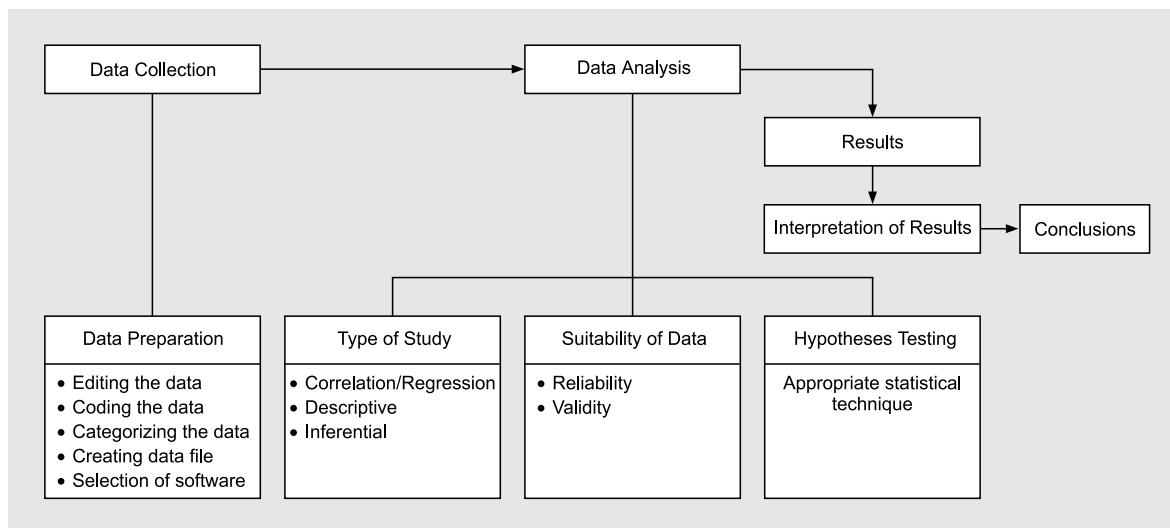
Their disadvantage is that the unit of measurement may not be the same as required by the users. For example, the size of a firm may be stated in terms of either number of employees, gross sales, gross profit, or total paid-up capital.

The scale of measurement may also be different from the one desired. For example, dividend declared by various companies may have breakup of 'less than 10 per cent' '10–15 per cent'; '15–20 per cent,' and so on. For a study requiring to know the number of companies who may have declared dividend of '16 per cent and above', such secondary data are of no use.

Robert W. Joselyn in his book *Designing and Marketing Research Project*, Petrocelli/Charter, 1977, New York, suggested an approach for evaluating the usefulness of secondary data and understanding their limitations. The flow chart showing the steps to be taken for evaluating the secondary data is shown in Fig. 1.6.

After data have been collected from a representative sample of the population, the next step is to analyse the data so that the research hypotheses can be tested. For this, some preliminary steps need to be followed. These steps help to prepare the data for analysis, ensure that the data obtained are reasonably good, and allow the results to be meaningfully interpreted. A flow diagram in Fig. 1.7 shows the data analysis process.

Figure 1.7
Flow Diagram of Data Analysis Process



Conceptual Questions 1B

30. A manager of a large corporation has recommended that a Rs 1000 raise be given to keep a valued subordinate from moving to another company. What internal and external sources of data might be used to decide whether such a salary increase is appropriate?
31. In the area of statistical measurement instruments such as questionnaires, *reliability* refers to the consistency of the measuring instrument and *validity* refers to the accuracy of the instrument. Thus, if a questionnaire yields comparable or similar results when completed by two equivalent groups of respondents, then the questionnaire can be described as being reliable. Does the fact that an instrument is reliable guarantee that it is also a valid instrument? Discuss.
32. Describe the three basic steps involved in the development and use of a written questionnaire prior to actual data analysis.
33. Describe the three general forms of questions that can be included in a questionnaire and give an example of each in the context of a political poll.
34. One can say that statistical inference includes an interest in statistical description as well, since the ultimate purpose of statistical inference is to *describe* population data. How then, does statistical inference differ from statistical description? Discuss.
35. In a recent study of causes of death in men 60 years of age and older, a sample of 120 men indicated that 48 died as a result of some form of heart disease.
 - (a) Develop a descriptive statistic that can be used as an estimate of the percentage of men 60 years of age or older who die from some form of heart disease
 - (b) Are the data on the causes of death qualitative or quantitative?
 - (c) Discuss the role of statistical inference in this type of medical research

- 36.** Determine whether each of the following random variables is categorical or numerical. If it is numerical, determine whether the phenomenon of interest is discrete or continuous.
- Amount of time the personal computer is used per week
 - Number of persons in the household who use the personal computer
 - Amount of money spent on clothing in the last month
 - Favourite shopping centre.
 - Amount of time spent shopping for clothing in the last month
- 37.** State whether each of the following variables is qualitative or quantitative and indicate the measurement scale that is appropriate for each.
- Age
 - Gender
 - Class rank
 - Make of automobile
 - Annual sales
 - Soft-drink size (small, medium, large)
 - Earnings per share
 - Method of payment (cash, check, credit card)
- 38.** A firm is interested in testing the advertising effectiveness of a new television commercial. As part of the test, the commercial is shown on a 6:30 p.m. local news programme in Delhi. Two days later, a market research firm conducts a telephone survey to obtain information on recall rates (percentage of viewers who recall seeing the commercial) and impressions of the commercial.
- What is the population for this study?
 - What is the sample for this study?
- 39.** Suppose the following information is obtained from a person on his application for a home loan from a bank:
- Place of residence: GK II, New Delhi
 - Type of residence: Single-family home
 - Date of birth: 14 August 1975
 - Monthly income: Rs 25,000
 - Occupation: Systems Engineer
 - Employer: Telecom company
 - Number of years at job: 5
 - Other income: Rs 30,000 per year
 - Marital status: Married
 - Number of children: 1
 - Loan requested: Rs 5,00,000
 - Term of Loan: 10 years
 - Other loan: Car
 - Amount of other loan: Rs 1,00,000
- Classify each of the 14 responses by type of data and level of measurement.
- 40.** Suppose that the Rotary Club was planning to survey 2000 of its members primarily to determine the percentage of its membership that currently own more than one car.
- Describe both the population and the sample of interest to the club
 - Describe the type of data that the club primarily wishes to collect
 - Develop the questionnaire needed by writing a series of five categorical questions and five numerical questions that you feel would be appropriate for this survey

It's not the figures themselves . . . , it's what you do with them that matters.

—K. A. C. Manderville

If you torture the data long enough, it will confess.

—Ronald Coase

Data Classification, Tabulation and Presentation

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- understand types of data and the basis of their classification.
- use techniques of organizing data in tabular and graphical form in order to enhance data analysis and interpretation.

2.1 INTRODUCTION

In Chapter 1, we learned how to collect data through primary and/or secondary sources. Whenever a set of data that we have collected contains a large number of observations, the best way to examine such data is to present it in some compact and orderly form. Such a need arises because data contained in a questionnaire are in a form which does not give any idea about the salient features of the problem under study. Such data are not directly suitable for *analysis* and *interpretation*. For this reason the data set is organized and summarized in such a way that patterns are revealed and are more easily interpreted. Such an arrangement of data is known as the *distribution* of the data. Distribution is important because it reveals the pattern of variation and helps in a better understanding of the phenomenon the data present.

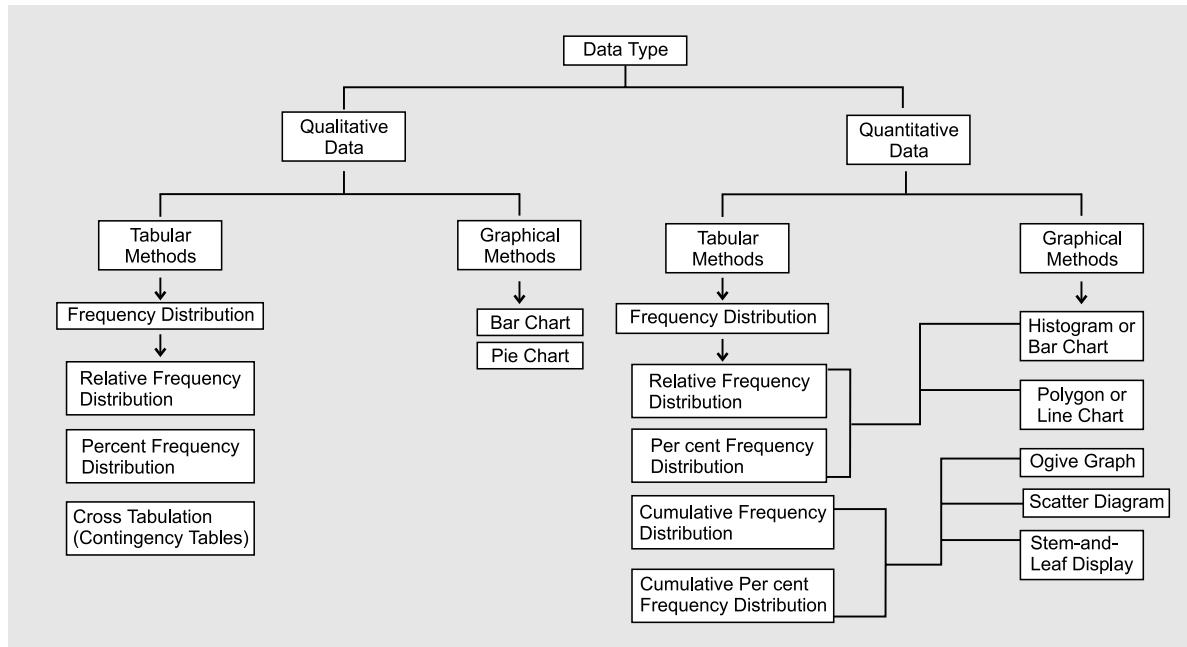
2.2 CLASSIFICATION OF DATA

Classification of data is the process of arranging data in groups/classes on the basis of certain properties. The classification of statistical data serves the following purposes:

- (i) It condenses the raw data into a form suitable for statistical analysis.
- (ii) It removes complexities and highlights the features of the data.
- (iii) It facilitates comparisons and drawing inferences from the data. For example, if university students in a particular course are divided according to sex, their results can be compared.

- (iv) It provides information about the mutual relationships among elements of a data set. For example, based on literacy and criminal tendency of a group of people, it can be established whether literacy has any impact or not on criminal tendency.
- (v) It helps in statistical analysis by separating elements of the data set into homogeneous groups and hence brings out the points of similarity and dissimilarity.

Figure 2.1
Tabular and Graphical Methods For
Summarizing Data



2.2.1 Requisites of Ideal Classification

The classification of data is decided after taking into consideration the nature, scope, and purpose of the investigation. However, an ideal classification should have following characteristics:

It should be unambiguous It is necessary that the various classes should be so defined that there is no room for confusion. There must be only one class for each element of the data set. For example, if the population of the country is divided into two classes, say literates and illiterates, then an exhaustive definition of the terms used would be essential.

Classes should be exhaustive and mutually exclusive Each element of the data set must belong to a class. For this, an extra class can be created with the title 'others' so as to accommodate all the remaining elements of the data set.

Each class should be mutually exclusive so that each element must belong to only one class. For example, classification of students according to the age: below 25 years and more than 20 years, is not correct because students of age 20 to 25 may belong to both the classes.

It should be stable The classification of a data set into various classes must be done in such a manner that if each time an investigation is conducted, it remains unchanged and hence the results of one investigation may be compared with that of another. For example, classification of the country's population by a census survey based on occupation suffers from this defect because various occupations are defined in different ways in successive censuses and, as such, these figures are not strictly comparable.

It should be flexible A classification should be flexible so that suitable adjustments can be made in new situations and circumstances. However, flexibility does not mean instability. The data should be divided into few major classes which must be further subdivided. Ordinarily there would not be many changes in the major classes. Only small sub-classes

may need a change and the classification can thus retain the merit of stability and yet have flexibility.

The term stability does not mean rigidity of classes. The term is used in a relative sense. One-time classification can not remain stable forever. With change in time, some classes become obsolete and have to be dropped and fresh classes have to be added. The classification may be called ideal if it can adjust itself to these changes and yet retain its stability.

2.2.2 Basis of Classification

Statistical data are classified after taking into account the nature, scope, and purpose of an investigation. Generally, data are classified on the basis of the following four bases:

Geographical Classification In geographical classification, data are classified on the basis of geographical or locational differences such as—cities, districts, or villages between various elements of the data set. The following is an example of a geographical distribution

City	:	Mumbai	Kolkata	Delhi	Chennai
Population density	:	654	685	423	205

(per square km)

Such a classification is also known as *spatial classification*. Geographical classifications are generally listed in alphabetical order. Elements in the data set are also listed by the frequency size to emphasize the importance of various geographical regions as in ranking the metropolitan cities by population density. The first approach is followed in case of reference tables while the second approach is followed in the case of summary tables.

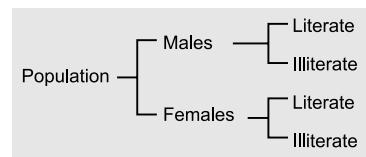
Chronological Classification When data are classified on the basis of time, the classification is known as chronological classification. Such classifications are also called *time series* because data are usually listed in chronological order starting with the earliest period. The following example would give an idea of chronological classification:

Year	:	1941	1951	1961	1971	1981	1991	2001
Population	:	31.9	36.9	43.9	54.7	75.6	85.9	98.6

(crore)

Qualitative Classification In qualitative classification, data are classified on the basis of descriptive characteristics or on the basis of attributes like sex, literacy, region, caste, or education, which cannot be quantified. This is done in two ways:

- (i) *Simple classification*: In this type of classification, each class is subdivided into two sub-classes and only one attribute is studied such as: male and female; blind and not blind, educated and uneducated, and so on.
- (ii) *Manifold classification*: In this type of classification, a class is subdivided into more than two sub-classes which may be sub-divided further. An example of this form of classification is shown in the box:



Quantitative Classification In this classification, data are classified on the basis of some characteristics which can be measured such as height, weight, income, expenditure, production, or sales.

Quantitative variables can be divided into the following two types. The term variable refers to any quantity or attribute whose value varies from one investigation to another.

- (i) *Continuous variable* is the one that can take any value within the range of numbers. Thus the height or weight of individuals can be of any value within the limits. In such a case, data are obtained by measurement.
- (ii) *Discrete (also called discontinuous) variable* is the one whose values change by steps or jumps and can not assume a fractional value. The number of children in a family, number of workers (or employees), number of students in a class, are few examples of a discrete variable. In such a case data are obtained by counting.

The following are examples of continuous and discrete variables in a data set:

Table 2.1

<i>Discrete Series</i>		<i>Continuous Series</i>	
<i>Number of Children</i>	<i>Number of Families</i>	<i>Weight (kg)</i>	<i>Number of Persons</i>
0	10	100 to 110	10
1	30	110 to 120	20
2	60	120 to 130	25
3	90	130 to 140	35
4	110	140 to 150	50
5	20		
	320		140

2.3 ORGANIZING DATA USING DATA ARRAY

The best way to examine a large set of numerical data is first to organize and present it in an appropriate tabular and graphical format.

Table 2.2 presents the total number of overtime hours worked for 30 consecutive weeks by machinists in a machine shop. The data displayed here are in *raw form*, that is, the numerical observations are not arranged in any particular order or sequence.

Table 2.2 Raw Data Pertaining to Total Time Hours Worked by Machinists

94	89	88	89	90	94	92	88	87	85
88	93	94	93	94	93	92	88	94	90
93	84	93	84	91	93	85	91	89	95

These raw data are not amenable even to a simple reading and do not highlight any characteristic/trend, such as the highest, the lowest, and the average weekly hours. Even a careful look at these data do not easily reveal any significant trend regarding the nature and pattern of variations therein. As such no meaningful inference can be drawn, unless these data are reorganized to make them more useful. For example, if we are to ascertain a value around which most of the overtime hours cluster, such a value is difficult to obtain from the raw data.

Moreover, as the number of observations gets large, it becomes more and more difficult to focus on the specific features in a set of data. Thus we need to organize the observation so that we can better understand the information that the data are revealing.

The raw data can be reorganized in a data array and frequency distribution. Such an arrangement enables us to see quickly some of the characteristics of the data we have collected.

When a raw data set is arranged in rank order, from the smallest to the largest observation or vice-versa, the ordered sequence obtained is called an *ordered array*. Table 2.3 reorganizes data given in Table 2.2 in the ascending order

Table 2.3 Ordered Array of Total Overtime Hours Worked by Machinists

84	84	85	85	87	88	88	88
88	89	89	89	90	90	91	91
92	92	93	93	93	93	93	93
94	94	94	94	94	95		

It may be observed that an ordered array does not summarize the data in any way as the number of observations in the array remains the same. However, a few advantages of ordered arrays are as under:

Advantages and Disadvantages of Ordered Array

Advantages The following are a few advantages of an ordered array:

- (i) It provides a quick look at the highest and lowest observations in the data within which individual values vary.
- (ii) It helps in dividing the data into various sections or parts.
- (iii) It enables us to know the degree of concentration around a particular observation.
- (iv) It helps to identify whether any values appear more than once in the array.

Disadvantages In spite of various advantages of converting a set of raw data into an ordered array, an array is a cumbersome form of presentation which is tiresome to construct. It neither summarizes nor organizes the data to present them in a more meaningful way. It also fails to highlight the salient characteristics of the data which may be crucial in terms of their relevance to decision-making.

The above task cannot be accomplished unless the observations are appropriately condensed. The best way to do so is to display them into a convenient number of groupings with the number of observations falling in different groups indicated against each. Such tabular summary presentation showing the number (frequency) of observations in each of several non-overlapping classes or groups is known as *frequency distribution* (also referred to as *grouped data*).

2.3.1 Frequency Distribution

A **frequency distribution** divides observations in the data set into conveniently established, numerically ordered classes (groups or categories). The number of observations in each class is referred to as *frequency* denoted as *f*.

Few examples of instances where frequency distributions would be useful are when (i) a marketing manager wants to know how many units (and what proportions or percentage) of each product sells in a particular region during a given period, (ii) a tax consultant desires to keep count of the number of times different size of firms are audited, and (iii) a financial analyst wants to keep track of the number of times the shares of manufacturing and service companies to be or gain order a period of time.

Frequency distribution: A tabular summary of data showing the number (frequency) of observations in each of several non-overlapping class intervals.

Advantages and Disadvantages of Frequency Distribution

Advantages The following are a few advantages of grouping and summarizing raw data in this compact form:

- (i) The data are expressed in a more compact form. One can get a deeper insight into the salient characteristics of the data at the very first glance.
- (ii) One can quickly note the pattern of distribution of observations falling in various classes.
- (iii) It permits the use of more complex statistical techniques which help reveal certain other obscure and hidden characteristics of the data.

Disadvantages A frequency distribution suffers from some disadvantages as stated below:

- (i) In the process of grouping, individual observations lose their identity. It becomes difficult to notice how the observations contained in each class are distributed. This applies more to a frequency distribution which uses the tally method in its construction.
- (ii) A serious limitation inherent in this kind of grouping is that there will be too much clustering of observations in various classes in case the number of classes is too small. This will cause some of the essential information to remain unexposed.

Hence, it is important that summarizing data should not be at the cost of losing essential details. The purpose should be to seek an appropriate compromise between having too much of details or too little. To be able to achieve this compromise, certain criteria are discussed for constructing a frequency distribution.

The frequency distribution of the number of hours of overtime given in Table 2.2 is shown in Table 2.4.

Table 2.4 Array and Tallies

Number of Overtime Hours	Tally	Number of Weeks (Frequency)
84		2
85		2
86	—	0
87		1
88		4
89		3
90		2
91		2
92		2
93		6
94		5
95		1
		30

Constructing a Frequency Distribution As the number of observations obtained gets larger, the method discussed above to condense the data becomes quite difficult and time-consuming. Thus to further condense the data into frequency distribution tables, the following steps should be taken:

- (i) Select an appropriate number of non-overlapping class intervals
- (ii) Determine the width of the class intervals
- (iii) Determine class limits (or boundaries) for each class interval to avoid overlapping.

1. Decide the number of class intervals The decision on the number of class groupings depends largely on the judgment of the individual investigator and/or the range that will be used to group the data, although there are certain guidelines that can be used. As a general rule, a frequency distribution should have at least five class intervals (groups), but not more than fifteen. The following two rules are often used to decide approximate number of classes in a frequency distribution:

- (i) If k represents the number of classes and N the total number of observations, then the value of k will be the smallest exponent of the number 2, so that $2^k \geq N$.

In Table 2.3 we have $N = 30$ observations. If we apply this rule, then we shall have

$$2^3 = 8 (< 30)$$

$$2^4 = 16 (< 30)$$

$$2^5 = 32 (> 30)$$

Thus we may choose $k = 5$ as the number of classes.

- (ii) According to Sturge's rule, the number of classes can be determined by the formula

$$k = 1 + 3.222 \log_e N$$

where k is the number of classes and $\log_e N$ is the logarithm of the total number of observations.

Applying this rule to the data given in Table 2.3, we get

$$\begin{aligned} k &= 1 + 3.222 \log 30 \\ &= 1 + 3.222 (1.4771) = 5.759 \approx 5 \end{aligned}$$

2. Determine the width of class intervals When constructing the frequency distribution it is desirable that the width of each class interval should be equal in size. The size (or width) of each class interval can be determined by first taking the difference between the largest and smallest numerical values in the data set and then dividing it by the number of class intervals desired.

$$\text{Width of class interval } (h) = \frac{\text{Largest numerical value} - \text{Smallest numerical value}}{\text{Number of classes desired}}$$

The value obtained from this formula can be rounded off to a more convenient value based on the investigator's preference.

From the ordered array in Table 2.3, the range is: $95 - 84 = 11$ hours. Using the above formula with 5 classes desired, the width of the class intervals is approximated as:

$$\text{Width of class interval} = \frac{11}{5} = 2.2 \text{ hours}$$

For convenience, the selected width (or interval) of each class is rounded to 3 hours.

3. Determine Class Limits (Boundaries) The limits of each class interval should be clearly defined so that each observation (element) of the data set belongs to one and only one class.

Each class has two limits—a *lower limit* and an *upper limit*. The usual practice is to let the lower limit of the first class be a convenient number slightly below or equal to the lowest value in the data set. In Table 2.3, we may take the lower class limit of the first class as 82 and the upper class limit as 85. Thus the class would be written as 82–85. This class interval includes all overtime hours ranging from 82 upto but not including 85 hours. The various other classes can be written as:

Overtime Hours (Class intervals)	Tallies	Frequency
82 but less than 85		2
85 but less than 88		3
88 but less than 91		9
91 but less than 94		10
94 but less than 97		6
		30

Mid-point of Class Intervals The main advantage of using the above summary table is that the major data characteristics become clear to the decision-maker. However, it is difficult to know how the individual values are distributed within a particular class interval without access to the original data. The **class mid-point** is the point halfway between the boundaries (both upper and lower class limits) of each class and is representative of all the observations contained in that class.

Arriving at the correct class mid-points is important, for these are used as representative of all the observations contained in their respective class while computing many important statistical measures. A mid-point is obtained by dividing the sum of the upper and lower class limits by two. Problems in computing mid-points arise when the class limits are ambiguous and not clearly defined.

The width of the class interval should, as far as possible, be equal for all the classes. If this is not possible to maintain, the interpretation of the distribution becomes difficult. For example, it will be difficult to say whether the difference between the frequencies of the two classes is due to difference in the concentration of observations in the two classes or due to the width of the class intervals being different.

Further, to facilitate computation of the summary measures discussed in Chapter 3, the width of the class intervals should preferably be not only the same throughout, it should also be a convenient number such as 5, 10, or 15. A width given by integers 7, 13, or 19 should be avoided.

Class mid-point: The point in each class that is halfway between the lower and upper class limits.

2.3.2 Methods of Data Classification

There are two ways in which observations in the data set are classified on the basis of class intervals, namely

- (i) Exclusive method, and
- (ii) Inclusive method

Exclusive Method When the data are classified in such a way that the upper limit of a class interval is the lower limit of the succeeding class interval (i.e. no data point falls into more than one class interval), then it is said to be the exclusive method of classifying data. This method is illustrated in Table 2.5.

Table 2.5 Exclusive Method of Data Classification

<i>Dividends Declared in per cent (Class Intervals)</i>	<i>Number of Companies (Frequencies)</i>
0–10	5
10–20	7
20–30	15
30–40	10

Such classification ensures continuity of data because the upper limit of one class is the lower limit of succeeding class. As shown in Table 2.5, 5 companies declared dividends ranging from 0 to 10 per cent, this means a company which declared exactly 10 per cent dividend would not be included in the class 0–10 but would be included in the next class 10–20. Since this point is not always clear, therefore to avoid confusion data are displayed in a slightly different manner, as given in Table 2.6.

Table 2.6

<i>Dividends Declared in per cent (Class Intervals)</i>	<i>Number of Companies (Frequencies)</i>
0 but less than 10	5
10 but less than 20	7
20 but less than 30	15
30 but less than 40	10

Inclusive Method When the data are classified in such a way that both lower and upper limits of a class interval are included in the interval itself, then it is said to be the inclusive method of classifying data. This method is shown in Table 2.7.

Table 2.7 Inclusive Method of Data Classification

<i>Number of Accidents (Class Intervals)</i>	<i>Number of Weeks (Frequencies)</i>
0–4	5
5–9	22
10–14	13
15–19	8
20–24	2

Remarks: 1. An exclusive method should be used to classify a set of data involving continuous variables and an inclusive method should be used to classify a set of data involving discrete variables.

2. If a continuous variable is classified according to the inclusive method, then certain adjustment in the class interval is needed to obtain continuity as shown in Table 2.8.

Table 2.8

<i>Class Intervals</i>	<i>Frequency</i>
30–44	28
45–59	32
60–74	45
75–89	50
90–104	35

To ensure continuity, first calculate correction factor as:

$$x = \frac{\text{Upper limit of a class} - \text{Lower limit of the next higher class}}{2}$$

and then subtract it from the lower limits of all the classes and add it to the upper limits of all the classes.

From Table 2.8, we have $x = (45 - 44) \div 2 = 0.5$. Subtract 0.5 from the lower limits of all the classes and add 0.5 to the upper limits. The adjusted classes would then be as shown in Table 2.9.

Table 2.9

<i>Class Intervals</i>	<i>Frequency</i>
29.5–44.5	28
44.5–59.5	32
59.5–74.5	45
74.5–89.5	50
89.5–104.5	35

3. Class intervals should be of equal size to make meaningful comparison between classes. In a few cases, extreme values in the data set may require the inclusion of *open-ended classes* and this distribution is known as an *open-ended distribution*. Such open-ended classes do not pose any problem in data analysis as long as only a few frequencies (or values) lie in these classes. However, an open-ended distribution is not fit for further mathematical calculations because *mid-value* which is used to represent the class, cannot be determined for an open-ended class. An example of an open-ended distribution is given in Table 2.10.

Table 2.10

<i>Age (Years)</i>		<i>Population (Millions)</i>
Under	5	17.8
5–17		44.7
18–24		29.9
25–44		69.6
45–64		44.6
65 and above		27.4
		234.0

Table 2.11 provides a tentative guide to determine an adequate number of classes.

Table 2.11 Guide to Determine the Number of Classes to Use

<i>Number of Observations, N</i>	<i>Suggested Number of Classes</i>
20	5
50	7
100	8
200	9
500	10
1000	11

Example 2.1: The following set of numbers represents mutual fund prices reported at the end of a week for selected 40 nationally sold funds.

10	17	15	22	11	16	19	24	29	18
25	26	32	14	17	20	23	27	30	12
15	18	24	36	18	15	21	28	33	38
34	13	10	16	20	22	29	29	23	31

Arrange these prices into a frequency distribution having a suitable number of classes.

Solution: Since the number of observations are 40, it seems reasonable to choose 6 ($2^6 > 40$) class intervals to summarize values in the data set. Again, since the smallest value is 10 and the largest is 38, therefore the class interval is given by

$$h = \frac{\text{Range}}{\text{Number of classes}} = \frac{38 - 10}{6} = \frac{28}{6} = 4.66 \approx 5$$

Now performing the actual tally and counting the number of values in each class, we get the frequency distribution by exclusive method as shown in Table 2.12:

Table 2.12: Frequency Distribution

Class Interval (Mutual Fund Prices, Rs)	Tally	Frequency (Number of Mutual Funds)
10 – 15		6
15 – 20		11
20 – 25		9
25 – 30		7
30 – 35		5
35 – 40		2
		40

Example 2.2: The take-home salary (in Rs) of 40 unskilled workers from a company for a particular month was.

2482	2392	2499	2412	2440	2444
2446	2540	2394	2365	2412	2458
2482	2394	2450	2444	2440	2494
2460	2425	2500	2390	2414	2365
2390	2460	2422	2500	2470	2428

Construct a frequency distribution having a suitable number of classes.

Solution: Since the number of observations are 30, we choose 5 ($2^5 > 30$) class intervals to summarize values in the data set. In the data set the smallest value is 2365 and the largest is 2500, so the width of each class interval will be

$$h = \frac{\text{Range}}{\text{Number of classes}} = \frac{2540 - 2365}{5} = \frac{175}{5} = 35$$

Sorting the data values into classes and counting the number of values in each class, we get the frequency distribution by exclusive method as

Table 2.13 Frequency Distribution

Class Interval (Salary, Rs)	Tally	Frequency (Number of Workers)
2365 – 2400		6
2400 – 2435		7
2435 – 2470		10
2470 – 2505		6
2505 – 2540		1
		30

Example 2.3: A computer company received a rush order for as many home computers as could be shipped during a six-week period. Company records provide the following daily shipments:

22	65	65	67	55	50	65
77	73	30	62	54	48	65
79	60	63	45	51	68	79
83	33	41	49	28	55	61
65	75	55	75	39	87	45
50	66	65	59	25	35	53

Group these daily shipments figures into a frequency distribution having the suitable number of classes.

Solution: Since the number of observations are 42, it seems reasonable to choose 6 ($2^6 > 42$) classes. Again, since the smallest value is 22 and the largest is 87, therefore the class interval is given by

$$h = \frac{\text{Range}}{\text{Number of classes}} = \frac{87 - 22}{6} = \frac{65}{6} = 10.833 \text{ or } 11$$

Now performing the actual tally and counting the number of values in each class, we get the following frequency distribution by inclusive method as shown in Table 2.14.

Table 2.14 Frequency Distribution

Class Interval (Number of Computers)	Tally	Frequency (Number of Days)
22 – 32		4
33 – 43		4
44 – 54		9
55 – 65		14
66 – 76		6
77 – 87		5
		42

Example 2.4: Following is the increase of D.A. in the salaries of employees of a firm at the following rates.

Rs 250 for the salary range up to Rs 4749

Rs 260 for the salary range from Rs 4750

Rs 270 for the salary range from Rs 4950

Rs 280 for the salary range from Rs 5150

Rs 290 for the salary range from Rs 5350

No increase of D.A. for salary of Rs 5500 or more. What will be the additional amount required to be paid by the firm in a year which has 32 employees with the following salaries (in Rs)?

5422	4714	5182	5342	4835	4719	5234	5035
5085	5482	4673	5335	4888	4769	5092	4735
5542	5058	4730	4930	4978	4822	4686	4730
5429	5545	5345	5250	5375	5542	5585	4749

Solution: Performing the actual tally and counting the number of employees in each salary range (or class), we get the following frequency distribution as shown in Table 2.15.

Table 2.15 Frequency Distribution

<i>Class Interval (Pay Range)</i>	<i>Tally</i>	<i>Frequency, f (Number of Employees)</i>	<i>Rate of D.A. (Rs x)</i>	<i>Total Amount Paid (Rs f x)</i>
upto 4749		8	250	2000
4750 – 4949		5	260	1300
4950 – 5149		5	270	1350
5150 – 5349		6	280	1680
5350 – 5549		8	290	2320
		<u>32</u>		<u>8650</u>

Hence additional amount required by the firm for payment of D.A. is Rs 8650.

Example 2.5: Following are the number of items of similar type produced in a factory during the last 50 days

21	22	17	23	27	15	16	22	15	23
24	25	36	19	14	21	24	25	14	18
20	31	22	19	18	20	21	20	36	18
21	20	31	22	19	18	20	20	24	35
25	26	19	32	22	26	25	26	27	22

Arrange these observations into a frequency distribution with both inclusive and exclusive class intervals choosing a suitable number of classes.

Solution: Since the number of observations are 50, it seems reasonable to choose 6 ($2^6 > 50$) or less classes. Since smallest value is 14, and the largest is 36 therefore the class interval is given by

$$h = \frac{\text{Range}}{\text{Number of classes}} = \frac{36 - 14}{6} = \frac{22}{6} = 3.66 \text{ or } 4$$

Performing the actual tally and counting the number of observations in each class, we get the following frequency distribution with inclusive class intervals as shown in Table 2.16.

Table 2.16 Frequency Distribution with Inclusive Class Intervals

<i>Class Intervals</i>	<i>Tally</i>	<i>Frequency (Number of Items Produced)</i>
14 – 17		6
18 – 21		18
22 – 25		15
26 – 29		5
30 – 33		3
34 – 33		3
		<u>50</u>

Converting the class intervals shown in Table 2.16 into exclusive class intervals is shown in Table 2.17.

Table 2.17 Frequency Distribution with Exclusive Class Intervals

<i>Class Intervals</i>	<i>Mid-Value of Class Intervals</i>	<i>Frequency (Number of Items Produced)</i>
13.5 – 17.5	15.5	6
17.5 – 21.5	19.5	18
21.5 – 25.5	23.5	15
25.5 – 29.5	27.5	5
29.5 – 33.5	31.5	3
33.5 – 37.5	34.5	3

2.3.3 Bivariate Frequency Distribution

The frequency distributions discussed so far involved only one variable and therefore called *univariate frequency distributions*. In case the data involve two variables (such as profit and expenditure on advertisements of a group of companies, income and expenditure of a group of individuals, supply and demand of a commodity, etc.), then frequency distribution so obtained as a result of cross classification is called *bivariate frequency distribution*. It can be summarized in the form of a *two-way (bivariate) frequency table* and the values of each variable are grouped into various classes (not necessarily same for each variable) in the same way as for univariate distributions.

If the data corresponding to one variable, say x , is grouped into m classes and the data corresponding to another variable, say y , is grouped into n classes, then bivariate frequency table will have $m \times n$ cells.

Frequency distribution of variable x for a given value of y is obtained by the values of x and vice-versa. Such frequencies in each cell are called *conditional frequencies*. The frequencies of the values of variables x and y together with their frequency totals are called the *marginal frequencies*.

Example 2.6: The following figures indicate income (x) and percentage expenditure on food (y) of 25 families. Construct a bivariate frequency table classifying x into intervals $200 - 300, 300 - 400, \dots$ and y into $10 - 15, 15 - 20, \dots$

Write the marginal distribution of x and y and the conditional distribution of x when y lies between 15 and 20.

x	y								
550	12	225	25	680	13	202	29	689	11
623	14	310	26	300	25	255	27	523	12
310	18	640	20	425	16	492	18	317	18
420	16	512	18	555	15	587	21	384	17
600	15	690	12	325	23	643	19	400	19

Solution: The two-way frequency table showing income (in Rs) and percentage expenditure on food is shown in Table 2.18.

Table 2.18

<i>Expenditure (y) (Percentage)</i>	<i>Income (x)</i>					<i>Marginal Frequencies, f_y</i>
	$200-300$	$300-400$	$400-500$	$500-600$	$600-700$	
10 – 15				(2)	(4)	6
15 – 20		(3)	(4)	(2)	(2)	11
20 – 25		(1)		(1)	(1)	3
25 – 30	(3)	(2)				5
<i>Marginal Frequencies, f_x</i>	3	6	4	5	7	25

The conditional distribution of x when y lies between 15 and 20 per cent is as follows:

Income (x) :	$200-300$	$300-400$	$400-500$	$500-600$	$600-700$
15%–20% :	0	3	4	2	2

Example 2.7: The following data give the points scored in a tennis match by two players X and Y at the end of twenty games:

(10, 12) (7, 11) (7, 9) (15, 19) (17, 21) (12, 8) (16, 10) (14, 14) (22, 18) (16, 7)

(15, 16) (22, 20) (19, 15) (7, 18) (11, 11) (12, 18) (10, 10) (5, 13) (11, 7) (10, 10)

Taking class intervals as: 5–9, 10–14, 15–19 ..., for both X and Y, construct

(i) Bivariate frequency table.

(ii) Conditional frequency distribution for Y given $X > 15$.

Solution: (i) The two-way frequency distribution is shown in Table 2.19.

Table 2.19 Bivariate Frequency Table

Player Y	Player X				Marginal Frequencies, f_y
	5–9	10–14	15–19	20–24	
5–9	(1)	(2)	(1)	—	4
10–14	(2)	(5)	(1)	—	
15–19	(1)	(1)	(3)	(1)	
20–24	—	—	(1)	(1)	
Marginal Frequencies, f_x	4	8	6	2	20

(ii) Conditional frequency distribution for Y given $X > 15$.

Player Y	Player X	
	15–19	20–24
5–9	1	—
10–14	1	—
15–19	3	1
20–24	1	1
	6	2

2.3.4 Types of Frequency Distributions

Cumulative frequency distribution: The cumulative number of observations less than or equal to the upper class limit of each class.

Cumulative Frequency Distribution Sometimes it is preferable to present data in a **cumulative frequency (cf) distribution** or simply a distribution which shows the cumulative number of observations below the upper boundary (limit) of each class in the given frequency distribution. A cumulative frequency distribution is of two types: (i) *more than* type and (ii) *less than* type.

In a *less than* cumulative frequency distribution, the frequencies of each class interval are added successively from top to bottom and represent the cumulative number of observations less than or equal to the class frequency to which it relates. But in the *more than* cumulative frequency distribution, the frequencies of each class interval are added successively from bottom to top and represent the cumulative number of observations greater than or equal to the class frequency to which it relates.

The frequency distribution given in Table 2.20 illustrates the concept of cumulative frequency distribution:

Table 2.20 Cumulative Frequency Distribution

Number of Accidents	Number of Weeks (Frequency)	Cumulative Frequency (less than)	Cumulative Frequency (more than)
0–4	5	5	45 + 5 = 50
5–9	22	5 + 22 = 27	23 + 22 = 45
10–14	13	27 + 13 = 40	10 + 13 = 23
15–19	8	40 + 8 = 48	2 + 8 = 10
20–24	2	48 + 2 = 50	2

From Table 2.20 it may be noted that cumulative frequencies are corresponding to the lower limit and upper limit of class intervals. The ‘less than’ cumulative frequencies are corresponding to the upper limit of class intervals and ‘more than’ cumulative frequencies are corresponding to the lower limit of class intervals shown in Table 2.21(a) and (b).

Table 2.21(a)

<i>Upper Limits</i>	<i>Cumulative Frequency (less than)</i>
less than 4	5
less than 9	27
less than 14	40
less than 19	48
less than 24	50

Table 2.21(b)

<i>Lower Limits</i>	<i>Cumulative Frequency (more than)</i>
0 and more	50
5 and more	45
10 and more	23
15 and more	10
20 and more	2

Relative Frequency Distribution To enrich data analysis it is sometimes important to show what percentage of observations fall within each class of a distribution instead of showing the actual class frequencies. To convert a frequency distribution into a corresponding **relative frequency distribution**, we divide each class frequency by the total number of observations in the entire distribution. Each relative frequency is thus a proportion as shown in Table 2.22.

Cumulative relative frequency distribution:

The cumulative number of observations less than or equal to the upper class limit of each class.

Table 2.22 Relative and Percentage Frequency Distributions

<i>Number of Accidents</i>	<i>Number of Weeks (Frequency)</i>	<i>Relative Frequency</i>	<i>Percentage Frequency</i>
0–4	5	$\frac{5}{50} = 0.10$	$\frac{5}{50} \times 100 = 10$
5–9	22	$\frac{22}{50} = 0.44$	$\frac{22}{50} \times 100 = 44$
10–14	13	$\frac{13}{50} = 0.26$	$\frac{13}{50} \times 100 = 26$
15–19	8	$\frac{8}{50} = 0.16$	$\frac{8}{50} \times 100 = 16$
20–24	2	$\frac{2}{50} = 0.04$	$\frac{2}{50} \times 100 = 4$
	<u>50</u>	<u>1.00</u>	<u>100</u>

Percentage Frequency Distribution A **percentage frequency distribution** is one in which the number of observations for each class interval is converted into a percentage frequency by dividing it by the total number of observations in the entire distribution. The quotient so obtained is then multiplied by 100, as shown in Table 2.22.

Cumulative percentage frequency distribution:

The cumulative percentage of observations less than or equal to the upper class limit of each class.

Example 2.8: Following are the number of two wheelers sold by a dealer during eight weeks of six working days each.

13	19	22	14	13	16	19	21
23	11	27	25	17	17	13	20
23	17	26	20	24	15	20	21
23	17	29	17	19	14	20	20
10	22	18	25	16	23	19	20
21	17	18	24	21	20	19	26

- (a) Group these figures into a table having the classes 10–12, 13–15, 16–18, . . . , and 28–30.
- (b) Convert the distribution of part (i) into a corresponding percentage frequency distribution and also a percentage cumulative frequency distribution.

Solution: (a) Frequency distribution of the given data is shown in Table 2.23.

Table 2.23 Frequency Distribution

<i>Number of Automobiles Sold (Class Intervals)</i>	<i>Tally</i>	<i>Number of Days (Frequency)</i>
10 – 12		2
13 – 15		6
16 – 18		10
19 – 21		16
22 – 24		8
25 – 27		5
28 – 30		1
		48

(ii) **Table 2.24: Percentage and More Than Cumulative Percentage Distribution**

<i>Number of Automobiles Sold (Class Intervals)</i>	<i>Number of Days (Frequency)</i>	<i>Cumulative Frequency</i>	<i>Percentage Frequency</i>	<i>Percentage Cumulative Frequency</i>
10 – 12	2	2	4.17	4.17
13 – 15	6	8	12.50	16.67
16 – 18	10	18	20.83	37.50
19 – 21	16	34	33.34	70.84
22 – 24	8	42	16.67	87.51
25 – 27	5	47	10.41	97.92
28 – 30	1	48	2.08	100.00
	48		100.00	

Conceptual Questions 2A

1. Explain the characteristics of a frequency distribution.
2. Illustrate two methods of classifying data in class-intervals.
3. Distinguish clearly between a continuous variable and a discrete variable. Give two examples of continuous variables and two examples of discrete variables that might be used by a statistician.
4. State whether the statement is true or false: The heights of rectangles erected on class intervals are proportional to the cumulative frequency of the class.
5. What is the basic property of virtually all data that lead to methods of describing and analysing data? How is a frequency distribution related to this property of data?
6. What are the advantages of using a frequency distribution to describe a body of raw data? What are the disadvantages?
7. When constructing a grouped frequency distribution, should equal intervals always be used? Under what circumstances should unequal intervals be used instead?
8. What are the advantages and disadvantages of using open-end intervals when constructing a group frequency distribution?
9. When constructing a group frequency distribution, is it necessary that the resulting distribution be symmetric? Explain.
10. Why is it necessary to summarize data? Explain the approaches available to summarize data distributions.
11. What are the objections to unequal class and open class intervals? State the conditions under which the use of unequal class intervals and open class intervals are desirable and necessary.
12. (a) What do you understand by cumulative frequency distribution?
 (b) What do you understand by bivariate or two-way frequency distribution?

Self-Practice Problems 2A

- 2.1** A portfolio contains 51 stocks whose prices are given below:

67	34	36	48	49	31	61	34
43	45	38	32	27	61	29	47
36	50	46	30	40	32	30	33
45	49	48	41	53	36	37	47
47	30	50	28	35	35	38	36
46	43	34	62	69	50	28	44
43	60	39					

Summarize these stock prices in the form of a frequency distribution.

- 2.2** Construct a frequency distribution of the data given below, where class interval is 4 and the mid-value of one of the classes is zero.

- 8	- 7	10	12	6	4	3	0	7
- 4	- 3	- 2	2	3	4	7	5	6
10	12	9	13	11	- 10	- 7	1	0
5	3	2	6	10	- 6	- 4		

- 2.3** Form a frequency distribution of the following data. Use an equal class interval of 4 where the lower limit of the first class is 10.

10	17	15	22	11	16	19	24	29
18	25	26	32	14	17	20	23	27
30	12	15	18	24	36	18	15	21
28	33	38	34	13	10	16	20	22
29	29	23	31					

- 2.4** If class mid-points in a frequency distribution of the ages of a group of persons are: 25, 32, 39, 46, 53, and 60, find:

- the size of the class-interval
- the class boundaries
- the class limits, assuming that the age quoted is the age completed on the last birthdays

- 2.5** The distribution of ages of 500 readers of a nationally distributed magazine is given below:

Age (in Years)	Number of Readers
Below 14	20
15–19	125
20–24	25
25–29	35
30–34	80
35–39	140
40–44	30
45 and above	45

Find the relative and cumulative frequency distributions for this distribution.

- 2.6** The distribution of inventory to sales ratio of 200 retail outlets is given below:

Inventory to Sales Ratio	Number of Retail Outlets
1.0–1.2	20
1.2–1.4	30
1.4–1.6	60
1.6–1.8	40
1.8–2.0	30
2.0–2.2	15
2.2–2.4	5

Find the relative and cumulative frequency distributions for this distribution.

- 2.7** A wholesaler's daily shipments of a particular item varied from 1,152 to 9,888 units per day. Indicate the limits of nine classes into which these shipments might be grouped.

- 2.8** A college book store groups the monetary value of its sales into a frequency distribution with the classes, Rs 400–500, Rs 501–600, and Rs 601 and over. Is it possible to determine from this distribution the amount of sales
- less than Rs 601
 - less than Rs 501
 - Rs 501 or more?

- 2.9** The class marks of distribution of the number of electric light bulbs replaced daily in an office building are 5, 10, 15, and 20. Find (a) the class boundaries and (b) class limits.

- 2.10** The marks obtained by 25 students in Statistics and Economics are given below. The first figure in the bracket indicates the marks in Statistics and the second in Economics.

(14, 12)	(0, 2)	(1, 5)	(7, 3)	(15, 9)
(2, 8)	(12, 18)	(9, 11)	(5, 3)	(17, 13)
(19, 18)	(11, 7)	(10, 13)	(13, 16)	(16, 14)
(6, 10)	(4, 1)	(9, 15)	(11, 14)	(8, 3)
(13, 11)	(14, 17)	(10, 10)	(11, 7)	(15, 15)

Prepare a two-way frequency table taking the width of each class interval as 4 marks, the first being less than 4.

- 2.11** Prepare a bivariate frequency distribution for the following data for 20 students:

Marks in Law:	10	11	10	11	11
	14	12	12	13	10
Marks in Statistics:	20	21	22	21	23
	23	22	21	24	23
Marks in Law:	13	12	11	12	10
	14	14	12	13	10
Marks in Statistics:	24	23	22	23	22
	22	24	20	24	23

Also prepare

- A marginal frequency table for marks in Law and Statistics
- A conditional frequency distribution for marks in Law when the marks in statistics are more than 22.

- 2.12** Classify the following data by taking class intervals such that their mid-values are 17, 22, 27, 32, and so on:

30	42	30	54	40	48	15	17	51
42	25	41	30	27	42	36	28	26
37	54	44	31	36	40	36	22	30
31	19	48	16	42	32	21	22	46
33	41	21						

[Madurai-Kamraj Univ., B.Com., 1995]

- 2.13** In degree colleges of a city, no teacher is less than 30 years or more than 60 years in age. Their cumulative frequencies are as follows:

Less than	:	60	55	50	45
		40	35	30	25
Total frequency :		980	925	810	675
		535	380	220	75

Find the frequencies in the class intervals 25–30, 30–35, ...

Hints and Answers

- 2.3** The classes for preparing frequency distribution by inclusive method will be

10–13, 14–17, 18–21, ..., 34–37, 38–41

- 2.4** (a) Size of the class interval = Difference between the mid-values of any two consecutive classes = 7

- (b) The class boundaries for different classes are obtained by adding (for upper class boundaries or limits) and subtracting (for lower class boundaries or limits) half the magnitude of the class interval that is, $7 \div 2 = 3.5$ from the mid-values.

Class Intervals : 21.5–28.5 28.5–35.5 35.5–42.5

Mid-Values : 25 32 39

Class Intervals : 42.5–49.5 49.5–56.5 56.5–63.5

Mid-Values : 46 53 60

- (c) The distribution can be expressed in inclusive class intervals with width of 7 as: 22–28, 29–35, ..., 56–63.

- 2.7** One possibility is 1000–1999, 2000–2999, 3000–3999, ... 9000–9999 units of the item.

Age (year)	Cumulative Frequency	Age	Frequency
Less than 25	75	20–25	75
Less than 30	220	25–30	220 – 75 = 145
Less than 35	380	30–35	380 – 220 = 160
Less than 40	535	35–40	535 – 380 = 155
Less than 45	675	40–45	675 – 535 = 140
Less than 50	810	45–50	810 – 675 = 135
Less than 55	925	50–55	925 – 810 = 115
Less than 60	980	55–60	980 – 925 = 55

2.4 TABULATION OF DATA

Meaning and Definition Tabulation is another way of summarizing and presenting the given data in a systematic form in rows and columns. Such presentation facilitates comparisons by bringing related information close to each other and helps in further statistical analysis and interpretation. Tabulation has been defined by two statisticians as:

- The logical listing of related quantitative data in vertical columns and horizontal rows of numbers with sufficient explanatory and qualifying words, phrases and statements in the form of titles, headings and explanatory notes to make clear the full meaning, context and the origin of the data. —Tuttle

This definition gives an idea of the broad structure of statistical tables and suggests that tabulation helps organize a set of data in an orderly manner to highlight its basic characteristics.

- Tables are means of recording, in permanent form, the analysis that is made through classification and by placing in a position just the things that are similar and should be compared. —Sechrist

This definition defines tabulation as the process of classifying the data in a systematic form which facilitates comparative studies of data sets.

2.4.1 Objectives of Tabulation

The above two definitions indicate that tabulation is meant to summarize data in a simplest possible form so that the same can be easily analysed and interpreted. A few objectives of tabulation defined by few statisticians are as follows:

- Tabulation is the process of condensing classified data in the form of a table so that it may be more easily understood, and so that any comparison involved may be more readily made.

—D. Gregory and H. Ward

- It is a medium of communication of great economy and effectiveness for which ordinary prose is inadequate. In addition to its formation in simple presentation, the statistical table is also a useful tool of analysis.

—D. W. Paden and E. F. Lindquist

The major objectives of tabulation are:

1. *To simplify the complex data:* Tabulation presents the data set in a systematic and concise form avoiding unnecessary details. The idea is to reduce the bulk of information (data) under investigation into a simplified and meaningful form.
2. *To economize space:* By condensing data into a meaningful form, space is saved without sacrificing on the quality and quantity of data.
3. *To depict trend:* Data condensed in the form of a table reveal the trend or pattern of data which otherwise cannot be understood in a descriptive form of presentation.
4. *To facilitate comparisons:* Data presented in a tabular form, having rows and columns, facilitate quick comparisons among its observations.
5. *To facilitate statistical comparisons:* Tabulation is a phase between classification of data and its presentation. Various statistical techniques such as measures of average and dispersion, correlation and regression, time series, and so on can be applied to analyse data and then interpret the results.
6. *To help as a reference:* When data are arranged in tables in a suitable form, they can easily be identified and can also be used as a reference for future needs.

2.4.2 Parts of a Table

Presenting data in a tabular form is an art. A statistical table should contain all the requisite information in a limited space but without any loss of clarity. There are variations in practice, but explained below are certain accepted rules for the construction of an ideal table:

1. **Table number:** A table should be numbered for easy identification and reference in future. The table number may be given either in the centre or side of the table but above the top of the title of the table. If the number of columns in a table is large, then these can also be numbered so that easy reference to these is possible.
2. **Title of the table:** Each table must have a brief, self-explanatory and complete title so that
 - (a) it should be able to indicate *nature* of data contained.
 - (b) it should be able to explain the *locality* (i.e. geographical or physical) of data covered.
 - (c) it should be able to indicate the *time* (or period) of data obtained.
 - (d) it should contain the *source* of the data to indicate the authority for the data, as a means of verification and as a reference. The source is always placed below the table.
3. **Caption and stubs:** The heading for columns and rows are called caption and stub, respectively. They must be clear and concise.

Two or more columns or rows with similar headings may be grouped under a common heading to avoid repetition. Such arrangements are called sub-captions or sub-stubs. Each row and column can also be numbered for reference and to facilitate comparisons. The caption should be written at the middle of the column in small letters to save space. If different columns are expressed in different units, then the units should be specified along with the captions.

The stubs are usually wider than column headings but must be kept narrow without sacrificing precision or clarity. When a stub occupies more than one line, the figures of the table should be written in the last line.

4. **Body:** The body of the table should contain the numerical information. The numerical information is arranged according to the descriptions given for each column and row.
5. **Prefactory or head note:** If needed, a prefactory note is given just below the title for its further description in a prominent type. It is usually enclosed in brackets and is about the unit of measurement.
6. **Footnotes:** Anything written below the table is called a footnote. It is written to further clarify either the title captions or stubs. For example if the data described in the table pertain to profits earned by a company, then the footnote may define whether it is profit before tax or after tax. There are various ways of identifying footnotes:
 - (a) Numbering footnotes consecutively with small number 1, 2, 3, ... or letters a, b, c ... or star *, **, ...
 - (b) Sometimes symbols like @ or \$ are also used to identify footnotes.

A blank model table is given below:

Table Number and Title [Head or Prefactory Note (if any)]

Stub Heading	Caption				Total (Rows)	
	Subhead		Subhead			
	Column-head	Column-head	Column-head	Column-head		
Stub Entries						
Total (Columns)						

Footnote :

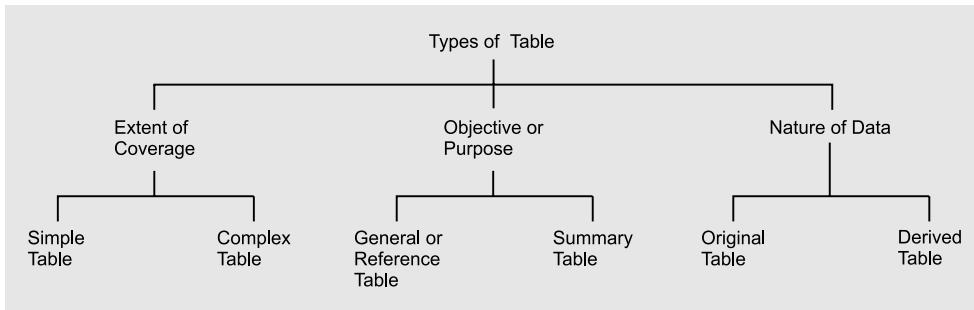
Source Note :

- Remarks:** 1. Information which is not available should be indicated by the letter N.A. or by dash (-) in the body of the table.
 2. Ditto marks ("), 'etc.' and use of the abbreviated forms should be avoided in the table.
 3. The requisites of a good statistical table given by various people are as flows:

- In the final analysis, there are only two rules in tabular presentation that should be applied rigidly. First, the use of common sense when planning a table, and second the viewing of the proposed table from the stand point of user. The details of mechanical arrangement must be governed by a single objective, that is, to make the statistical table as easy to read and to understand as the nature of the material will permit. —J. C. Capt
- A good statistical table is not a mere careless grouping of columns and rows of figures, it is a triumph of ingenuity and technique, a master-piece of economy of space combined with a maximum of clearly presented information. To prepare a first class table, one must have a clear idea of the facts to be presented, the contrasts to be stressed, the points upon which emphasis is to be placed and lastly a familiarity with the technique of preparation. —Harry Jerome
- In collection and tabulation, commonsense is the chief requisite and experience, the chief teacher. —A. L. Bowley

2.4.3 Types of Tables

The classification of tables depends on various aspects: objectives and scope of investigation, nature of data (primary or secondary) for investigation, extent of data coverage, and so on. The different types of tables used in statistical investigations are as follows:



Simple and Complex Tables In a *simple table* (also known as one-way table), data are presented based on only one characteristic. Table 2.25 illustrates the concept.

Table 2.25 Candidates Interviewed for Employment in a Company

Candidate's Profile	Number of Candidates
Experienced	50
Inexperienced	70
Total	120

The *complex table* also known as a manifold table is that in which data are presented according to two or more characteristics simultaneously. The complex tables are two-way or three-way tables according to whether two or three characteristics are presented simultaneously.

- (a) *Double or Two-Way Table:* In such a table, the variable under study is further subdivided into two groups according to two inter-related characteristics. For example, if the total number of candidates given in Table 2.24 are further divided according to their sex, the table would become a two-way table because it would reveal information about two characteristics namely, male and female. The new shape of the table is shown in Table 2.26.

Table 2.26 Candidates Interviewed for Employment in a Company

Candidates Profile	Number of Candidates		Total
	Males	Females	
Experienced	35	15	50
Inexperienced	10	60	70
Total	45	75	120

- (b) *Three-Way Table:* In such a table, the variable under study is divided according to three interrelated characteristics. For example, if the total number of males and females candidates given in Table 2.26 are further divided according to the marital status, the table would become a three-way. The new shape of the table is shown in Table 2.27.

Table 2.27 Candidates Interviewed for Employment in a Company

Candidates Profile	Number of Candidates						Total	
	Males			Females				
	Married	Unmarried	Total	Married	Unmarried	Total		
Experienced	15	20	35	5	10	15	50	
Inexperienced	2	8	10	10	50	60	70	
Total	17	28	45	15	60	75	120	

- (c) *Manifold (or Higher Order) Table:* Such tables provide information about a large number of inter-related characteristics in the data set. For example, if the data given in Table 2.27 is also available for other companies, then table would become a manifold table.

2.4.4 General and Summary Tables

General tables are also called *reference* or *repository tables*. In such a table, data are presented in detail so as to provide information for general or reference use on the same subject. Such tables are usually large in size and are generally given in the appendix for reference. Various people have identified the purpose of such tables which are given below:

- Primary and usually the sole purpose of a reference table is to present data in such a manner that individual items may be found readily by a reader.
—Croxton and Cowden
- Reference tables contain ungrouped data basic for a particular report, usually containing a large amount of data and frequently selected to a tabular appendix.
—Horace Secrist
- These tables are those in which data are recorded not the detailed data which have been analysed but rather the results of the analysis.
—John I. Griffin

Data published by various ministries, autonomous bodies, or institutions pertaining to employment, production, public expenditure, taxation, population, and so on are examples of such tables.

2.4.5 Original and Derived Tables

Original tables are also called *classification tables*. Such a table contains data collected from a primary source. But if the information given in a table has been derived from a general table, then such a table is called a *derived table*. For example, if from a general table, certain averages, ratios, or percentages are derived, then the table containing such information would be a derived table.

Example 2.9: A state government has taken up a scheme of providing drinking water to every village. During the first four years of a five-year plan, the government has installed 39,664 tubewells. Out of the funds earmarked for natural calamities, the government has sunk 14,072 tubewells during the first four years of the plan. Thus, out of the plan fund 9245 and 8630 tubewells were sunk, in 2000–2001 and 2001–2002, respectively. Out of the natural calamities fund, the number of tubewells sunk in 1998–99 and 1999–2000 were 4511 and 637, respectively. The expenditure for 2000–2001 and 2001–2002 was Rs 863.41 lakh and Rs 1185.65 lakh, respectively.

The number of tubewells installed in 2002–2003 was 16,740 out of which 4800 were installed out of the natural calamities fund and the expenditure of sinking of tubewells during 2002–2003 was Rs 1411.17 lakh.

The number of tubewells installed in 2003–2004 was 13,973, out of which 9849 tubewells were sunk out of the fund for the plan and the total expenditure during the first four years was Rs 5443.05 lakh.

Represent this data in a tabular form.

Solution: The data of the problem is summarized in Table 2.28.

Table 2.28 Tubewells for Drinking Water for Villages in a State

Year	Number of Tubewells		Expenditure (in Rs lakh)
	Out of Fund Plan	Out of Natural Calamities Fund	
2001–2001	9245	4511	863.41
2001–2002	8630	637	1185.65
2002–2003	(16,740 – 4800) = 11,940	4800	1411.17
2003–2004	9849	(13,973 – 9849) = 4124	1982.82
Total	39,664	14,072	5,443.05

Example 2.10: In a sample study about coffee-drinking habits in two towns, the following information was received:

Town A : Females were 40 per cent. Total coffee drinkers were 45 per cent and male non-coffee drinkers were 20 per cent

Town B : Males were 55 per cent. Male non-coffee drinkers were 30 per cent and female coffee drinkers were 15 per cent.

Represent this data in a tabular form.

Solution: The given data is summarized in Table 2.29.

Table 2.29 Coffee Drinking Habit of Towns A and B (in percentage)

Attribute	Town A			Town B			Total (1) + (2)
	Males	Females	Total (1)	Males	Females	Total (2)	
Coffee drinkers	(45 – 5) = 40	(40 – 35) = 5	45	(55 – 30) = 25	15	40	85
Non-coffee drinkers	20	(55 – 20) = 35	(100 – 45) = 55	30	(60 – 30) = 30	(100 – 40) = 60	115
Total	(100 – 40) = 60	40	100	55	(100 – 55) = 45	100	200

Example 2.11: Industrial finance in India has showed great variation in respect of sources of funds during the first, second, and third five-year plans. There were two main sources—internal and external. The internal sources of funds are—depreciation, free reserves and surplus. The external sources of funds are—capital issues, borrowings.

During the first plan, internal and external sources accounted for 62 per cent and 38 per cent of the total, and of the depreciation, fresh capital, and other sources formed 29 per cent, 7 per cent, and 10.6 per cent respectively.

During the second plan, internal sources decreased by 17.3 per cent compared to the first plan, and depreciation was 24.5 per cent. The external finance during the same period consisted of 10.9 per cent fresh capital and 28.9 per cent borrowings.

Compared to the second plan, external finance during the third plan decreased by 4.4 per cent, and borrowings and ‘other sources’ were 29.4 per cent and 14.9 per cent respectively. During the third plan, internal finance increased by 4.4 per cent and free reserves and surplus formed 18.6 per cent.

Tabulate this information with the above details as clearly as possible observing the rules of tabulation.

Solution: The given information is summarized in Table 2.30.

Table 2.30 Pattern of Industrial Finance (in Percentage)

Five Year Plan	Sources of Funds						
	Internal			External			
	Depreciation	Free Reserves and Surplus	Total	Capital Issues	Borrowings	Other Sources	Total
First	29	62 – 29 = 33	62	7	38 – 7 – 10.6 = 20.4	10.6	38
Second	24.5	44.7 – 24.5 = 20.2	62 – 17.3 = 44.7	10.9	28.9	55.3 – 10.9 – 28.9 = 15.5	100 – 44.7 = 55.3
Third	49.1 – 18.6 = 30.5	18.6	44.7 + 4.4 = 49.1	50.9 – 29.4 – 14.9 = 6.6	29.4	14.9	55.3 – 4.4 = 50.9

Example 2.12: The following information about weather conditions at different stations were recorded at 8.30 a.m. on Thursday, 29 August 1990.

At Ahmednagar station, the maximum and minimum temperature in 24 hrs were 28°C and 20°C respectively. The rainfall in the past 24 hrs at Ahmednagar was nil. Since 1 June the rainfall was 185 mm which is 105 mm below normal.

Bangalore's minimum and maximum temperatures in 24 hrs for the day were 19°C and 23°C respectively. It had no rainfall in the past 24 hrs and since 1 June the rainfall was 252 mm which is 54 mm below normal.

The minimum temperature at Udaipur was 21°C and the rainfall in the past 24 hrs was nil. Since 1 June it experienced 434 mm of rainfall which is 24 mm below normal.

Panagarh's maximum temperature in 24 hrs was 28°C. It had 4 mm of rain in the past 24 hrs and since 1 June it had 955 mm of rain.

Kolkata's maximum and minimum day temperatures were 30°C and 26°C respectively. It had 3 mm of rainfall in the past 24 hrs. Since 1 June it experienced a rainfall of 1079 mm which is 154 mm above normal.

Present the above data in a tabular form.

Solution: The given information is summarized in Table 2.31.

Table 2.31 Weather Conditions at Different Stations (at 8.30 a.m. on 29 August 1990)

Stations	In 24 Hours		Rainfall (mm)			
	Temperature (°C)		Past 24 hrs	Since June 1	Above Normal	Below Normal
	Min.	Max.			—	—
Ahmednagar	20	28	0	185	—	105
Bangalore	19	23	0	252	—	54
Udaipur	21	—	0	434	—	24
Panagarh	—	28	4	955	—	—
Kolkata	26	30	3	1079	154	—

Example 2.13: Present the following data in a tabular form:

A certain manufacturer produces three different products 1, 2, and 3. Product 1 can be manufactured in one of the three plants A, B, or C. However, product 2 can be manufactured in either plant B or C, whereas plant A or B can manufacture product 3. Plant A can manufacture in an hour 10 pieces of 1 or 20 pieces of 3, 20 pieces of 2, 15 pieces of 1, or 16 pieces of 3 can be manufactured per hour in plant B. C can produce 20 pieces of 1 or 18 pieces of 2 per hour.

Wage rates per hour are Rs 20 at A, Rs 40 at B and Rs 25 at C. The costs of running plants A, B, and C are respectively Rs 1000, 500, and 1250 per hour. Materials and other costs directly related to the production of one piece of the product are respectively Rs 10 for 1, Rs 12 for 2, and Rs 15 for 3. The company plans to market product 1 at Rs 15 per piece, product 2 at Rs 18 per piece and product 3 at Rs 20 per piece.

Solution: The given information is summarized in Table 2.32.

Table 2.32 Production Schedule of a Manufacturer

Plant	Rate of Manufacturing Per Hour (Pieces)			Wage Rates Per Hour (Rs)	Cost of Running Plant Per Hour (Rs)
	1	2	3		
A	10	—	20	20	1000
B	15	20	16	40	500
C	20	18	—	25	1250
Material and other direct costs per piece (Rs)	10	12	15		
Product price per piece (Rs)	15	18	20		

Example 2.14: Transforming the ratios into corresponding numbers, prepare a complete table for the following information. Give a suitable title to the table.

In the year 2000 the total strength of students of three colleges X, Y, and Z in a city were in the ratio 4 : 2 : 5. The strength of college Y was 2000. The proportion of girls and boys in all colleges was in the ratio 2 : 3. The faculty-wise distribution of boys and girls in the faculties of Arts, Science, and Commerce was in the ratio 1 : 2 : 2 in all the three colleges.

Solution: The data of the problem is summarized in Table 2.33.

Table 2.33 Distribution of Students According to Faculty and Colleges in the Year 2000

Colleges	Faculty									<i>Total</i> (1) + (2) + (3)	
	Arts			Science			Commerce				
	Boys	Girls	Total (1)	Boys	Girls	Total (2)	Boys	Girls	Total (3)		
X	480	320	800	960	640	1600	960	640	1600	4000	
Y	240	160	400	480	320	800	480	320	800	2000	
Z	600	400	1000	1200	800	2000	1200	800	2000	5000	
Total	1320	880	2200	2640	1760	4400	2640	1760	4400	11,000	

Example 2.15: Represent the following information in a suitable tabular form with proper rulings and headings:

The annual report of a Public Library reveals the following information regarding the reading habits of its members.

Out of the total of 3718 books issued to the members in the month of June, 2100 were fiction. There were 467 members of the library during the period and they were classified into five classes—A, B, C, D, and E. The number of members belonging to the first four classes were respectively 15, 176, 98, and 129, and the number of fiction books issued to them were 103, 1187, 647, and 58 respectively. The number of books, other than text books and fiction, issued to these four classes of members were respectively 4, 390, 217, and 341. Text books were issued only to members belonging to classes C, D, and E, and the number of text books issued to them were respectively 8, 317, and 160.

During the same period, 1246 periodicals were issued. These include 396 technical journals of which 36 were issued to member of class B, 45 to class D, and 315 to class E.

To members of classes B, C, D, and E the number of other journals issued were 419, 26, 231, and 99, respectively.

The report, however, showed an increase of 4.1 per cent in the number of books issued over last month, though there was a corresponding decrease of 6.1 per cent in the number of periodicals and journals issued to members.

Solution: The data of the problem is summarized in Table 2.35.

Table 2.35 Reading Habits of the Members of Public Library

	Type of Book Issued	Class of Members					Total for the Month	
		A	B	C	D	E	June	May
Books	Fiction	103	1187	647	58	105	2100	2018
	Textbooks	—	—	8	317	160	485	466
	Others	4	390	217	341	181	1133	1089
	Total	107	1577	872	716	446	3718	3573
Periodicals and Journals	Technical Journals	—	36	—	45	315	396	420
	Others	75	419	26	231	99	850	902
	Total	75	455	26	276	414	1246	1322

Note: The figures for the month of May were calculated on the basis of percentage changes for each type of reading material given in the text.

Conceptual Questions 2B

13. What is a statistical table? Explain clearly the essentials of a good table.
14. (a) What are the different components of a table?
 (b) What are the chief functions of tabulation? What precautions would you take in tabulating statistical data?
 (c) What are the characteristics of a good table?
15. Explain the role of tabulation in presenting business data, and discuss briefly the different methods of presentation.
16. Explain the terms 'classification' and 'tabulation'. Point out their importance in a statistical investigation. What precautions would you take in tabulating statistical data?
17. In classification and tabulation, common sense is the chief requisite and experience the chief teacher. Comment.
18. What are the requisites of a good table? State the rules that serve as a guide in tabulating statistical data.
19. Distinguish between classification and tabulation. Mention the requisites of a good statistical table.
20. Explain how you would tabulate statistics of deaths from sexual diseases in different states of India for a period of five years.
21. Explain the purpose of tabular presentation of statistical data. Draft a form of tabulation to show the distribution of population according to (i) community by age, (ii) literacy, (iii) sex, and (iv) marital status.

Self-Practice Problems 2B

- 2.14 Draw a blank table to show the number of candidates sex-wise appearing in the pre-university, first year, second year, and third year examinations of a university in the faculties of Arts, Science, and Commerce in a certain year.
- 2.15 Let the national income of a country for the years 2000–01 and 2001–02 at current prices be 80,650, 90,010, and 90,530 crore of rupees respectively, and per capita income for these years be 1050, 1056, and 1067 rupees. The corresponding figures of national income and per capita income at 1999–2000 prices for the above years were 80,650, 80,820, and 80,850 crore of rupees and 1050, 1051 and 1048 respectively. Present this data in a table.
- 2.16 Present the following information in a suitable form supplying the figure not directly given. In 2004, out of a total of 4000 workers in a factory, 3300 were members of a trade union. The number of women workers employed was 500 out of which 400 did not belong to any union.

In 2003, the number of workers in the union was 3450 of which 3200 were men. The number of non-union workers was 760 of which 330 were women.
- 2.17 Of the 1125 students studying in a college during a year, 720 were SC/ST, 628 were boys, and 440 were science students; the number of SC/ST boys was 392, that of boys studying science 205, and that of SC/ST students studying science 262; finally the number of science students among the SC/ST boys was 148. Enter these frequencies in a three-way table and complete the table by obtaining the frequencies of the remaining cells.
- 2.18 A survey of 370 students from the Commerce Faculty and 130 students from the Science Faculty revealed that 180 students were studying for only C.A. Examinations, 140 for only Costing Examinations, and 80 for both C.A. and Costing Examinations. The rest had opted for part-time Management Courses. Of those studying for

Costing only, 13 were girls and 90 boys belonged to the Commerce Faculty. Out of the 80 studying for both C.A. and Costing, 72 were from the Commerce Faculty amongst whom 70 were boys. Amongst those who opted for part-time Management Courses, 50 boys were from the Science Faculty and 30 boys and 10 girls from the Commerce Faculty. In all, there were 110 boys in the Science Faculty.

Present this information in a tabular form. Find the number of students from the Science Faculty studying for part-time Management Courses.

- 2.19 An Aluminium Company is in possession of certain scrap materials with known chemical composition. Scrap 1 contains 65 per cent aluminium, 20 per cent iron, 2 per cent copper, 2 per cent manganese, 3 per cent magnesium and 8 per cent silicon. The aluminium content of scrap 2, scrap 3, and scrap 4 are 70 per cent, 80 per cent and 75 per cent respectively. Scrap 2 contains 15 per cent iron, 3 per cent copper, 2 per cent manganese, 4 per cent magnesium and the rest silicon. Scrap 3 contains 5 per cent iron. The iron content of scrap 4 is the same as that of scrap 3, scrap 4 contains twice as much percentage of copper as scrap 3. scrap 3 contains 1 per cent copper. Scrap 3 contains manganese which is 3 times as much as copper it contains. The percentage of magnesium and silicon in scrap 3 are 3 per cent and 8 per cent respectively. The magnesium and silicon contents of scrap 4 are respectively 2 times and 3 times its manganese contents. The company also purchases some aluminium and silicon as needed. The aluminium purchased contains 96 per cent pure aluminium, 2 per cent iron, 1 per cent copper and 1 per cent silicon respectively, whereas the purchased silicon contains 98 per cent silicon and 2 per cent iron respectively. Present the above data in a table.
- 2.20 Present the following data in a suitable tabular form with appropriate headings:

A pilot survey carried out a few years before yielded the following estimates of livestock numbers and milk production in three regions, namely, Punjab Plains (PP), Punjab Hills (PH), and Eastern U.P. (EU) (only the estimates for the rural sector are quoted here.) The total number of cows was 4396, 2098 and 15,170 thousand, respectively in three regions, namely, PP, PH and EU, and the corresponding number for buffaloes were 4,092, 765 and 5,788 thousand. The percentages of animals producing milk were 47, 40 and 37 for cows in PP, PH, and EU, respectively, the corresponding figures for buffaloes being 58, 49 and 47. The average daily milk yield per animal was 2.52, 0.51, and 0.68 kg for cows in PP, PH, and EU respectively; and for buffaloes the yield figures were 4.10, 2.35 and 1.86 kg respectively.

- 2.21** Prepare a blank tabular layout with appropriate headings for presenting the estimates of the number of unemployed persons obtained from a sample survey covering three states namely, Bihar, Orissa, and West Bengal. The estimates should be presented separately for the three states, for rural and urban areas of each state, and also separately for persons in different levels of general education (illiterate, literate below primary, primary, secondary, graduate and above). The table should show the number of unemployed persons in each region and education level as well as the percentage of such persons to the corresponding total population. It should also present relevant sub-totals and totals.
- 2.22** A state was divided into three areas: administrative district, urban district, and rural district. A survey of housing conditions was carried out and the following information was gathered:

There were 67,71,000 buildings of which 17,61,000 were in rural district. Of the buildings in urban district 40,64,000 were inhabited and 45,000 were under construction. In the administrative district 40,000 buildings were uninhabited and 5000 were under construction of the total of 6,16,000. The total buildings in the city that are under construction are 62,000 and those uninhabited are 4,49,000. Tabulate this information.

- 2.23** Draw up a blank table to show the number of employees in a large commercial firm, classified according to (i) Sex: male and female; (ii) three age groups : below 30, 30 and above but below 45, 45 and above; and (iii) four income-groups: below Rs 400, Rs 400–750, Rs 750–1000, and above Rs 1000.
- 2.24** Transform the ratios into corresponding numbers to prepare a complete table for the following information. Give a suitable title to the table.

In the year 1997, the total strength of students of three colleges A, B, and C in a city was in the ratio

3 : 1 : 4. The strength of college B was 800. The proportion of girls and boys in all colleges was in the ratio 1 : 3. The faculty-wise distribution of girls and boys in the faculties of Arts, Science, and Commerce was in the ratio 2 : 1 : 2 in all the three colleges.

- 2.25** The ‘Financial Highlights’ of a public limited company in recent years were as follows:

In the year ending on 31 March 1998 the turnover of the company, including other income, was Rs 157 million. The profit of the company in the same year before tax, investment allowance, reserve, and prior year’s adjustment was Rs 19 million, and the profit after tax, investment allowance, reserve, and prior year’s adjustment was Rs 8 million. The dividend declared by the company in the same year was 20 per cent. The turnover, including other income, for the years ending on 31 March 1999, 2000, and 2001 were Rs 169, 191, and 197 million respectively. For the year ending on 31 March 1999 the profit before tax, investment allowance, reserve, and prior year’s adjustment was Rs 192 million and the profit after tax, and so on Rs. 7.5 million, while the dividend declared for the same year was 17 per cent. For the year ending on 31 March 2000, 2001, and 2002 the profits before tax, investment allowance, reserve, and prior year’s adjustment were Rs 21, 12, 13 million respectively, while the profits after tax, and so on, of the above three years were Rs 9.5, 4, and 9 million respectively. The turnover, including other income, for the year ending on 31 March 2002 was Rs 243 million. The dividend declared for the year ending on 31 March 2000–02 was 17 per cent, 10 per cent and 20 per cent respectively. Present the above data in a table.

- 2.26** Present the following information in suitable form:

In 1994, out of a total of 1950 workers of a factory, 1400 were members of a trade union.

The number of women employed was 400 of which 275 did not belong to a trade union. In 1999, the number of union workers increased to 1780 of which 1490 were men. On the other hand, the number of non-union workers fell to 408 of which 280 were men.

In the year 2004, there were 2000 employees who belonged to a trade union and 250 did not belong to a trade union. Of all the employees in 2000, 500 were women of whom only 208 did not belong to a trade union.

- 2.27** In a trip organized by a college, there were 50 persons, each of whom paid Rs 2500 on an average. There were 40 students each of whom paid Rs 2700. Members of the teaching staff were charged at a higher rate. The number of servants was 5 and they were not charged any amount. The number of ladies was 20 per cent of the total and one was a lady teacher. Tabulate the above information.

Hints and Answers

2.14 Distribution of candidates appearing in various university examinations

Faculty	Boys					Girls				
	Pre-Univ.	First year	Second year	Third year	Total	Pre-Univ.	First year	Second year	Third year	Total
Arts										
Science										
Commerce										
Total										

2.15 National income and per capita income of the country

For the year 1999-2000 to 2001-2002

Year	National Income			Per Capita Income	
	At Current Prices (Rs in crore)	At 1999–2000 Prices (Rs in crore)		At Current Prices	At 1999–2000 Prices
1999–2000	80,650	80,650		1050	1050
2000–2001	90,010	80,820		1056	1051
2001–2002	90,530	80,850		1067	1048

2.16 Members of union by sex

Year	2003			2004		
	Males	Females	Total	Males	Females	Total
Member	3300	250	3450	3200	100	3300
Non-member	430	330	760	300	400	700
Total	3630	580	4210	3500	500	4000

2.17 Distribution of College Students by Caste and Faculty

Faculty	Boys			Girls		
	SC/ST	Non-SC/ST	Total	SC/ST	Non-SC/ST	Total
Science	148	57	205	114	121	235
Arts	244	179	423	214	48	262
Total	392	236	628	328	169	497

2.18 Distribution of students according to Faculty and Professional Courses

Faculty Courses	Commerce			Science			Total		
	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total
Part-time Management	30	10	40	50	10	60	80	20	100
CA only	150	8	158	16	6	22	166	14	180
Costing only	90	10	100	37	3	40	127	13	140
CA and Costing	70	2	72	7	1	8	77	3	80
Total	340	30	370	110	20	130	450	50	500

2.19 The chemical composition of Scraps and Purchased Minerals

Materials	Chemical Composition (in percentage)					
	Aluminium	Iron	Copper	Manganese	Magnesium	Silicon
Scrap 1	65	20	2	2	3	8
Scrap 2	70	15	3	2	4	6
Scrap 3	80	5	1	3	3	8
Scrap 4	75	5	2	3	6	9
Aluminium	96	2	1	—	—	1
Silicon	—	2	—	—	—	98

2.25 Financial highlights of the Public Ltd.. Co.

Year Ended 31 March	Turnover Including Other Income (in Million of Rs)	Profit Before Tax, Investment Allowance Reserve and Prior Year Adjustment (in Million of Rs)	Profit After Tax Investment Allowance Reserve and Prior Year Adjustment (in Million of Rs)	Per cent
1998	157	19	8	20
1999	169	18	7.5	17
2000	191	21	9.5	17
2001	197	12	4	10
2002	243	13	9	20

2.26 Trade-union membership

Category	1994			1999			2004		
	Member	Non-member	Total	Member	Non-member	Total	Member	Non-member	Total
Men	1,275	275	1550	1490	280	1770	1708	42	1750
Women	125	275	400	290	128	418	292	208	500
Total	1400	550	1950	1780	408	2188	2000	250	2250

2.27	Type of Participants	Sex			Contribution	
		Males	Females	Total	Per Member (Rs)	Total Contribution (Rs)
	Students	31	9	40	2700	1,08,000
	Teaching staff	4	1	5	3400	17,000
	Servant	5	—	5	—	—
	Total	40	10	50	—	1,25,000

Notes : 1. Total contribution = Average contribution × Number of persons in the group
 $= 2500 \times 50 = \text{Rs } 1,25,000$

2. Per head contribution of teaching staff = $\frac{\text{Total contribution} - \text{Contribution of students}}{\text{Number of teaching staff}}$
 $= \frac{1,25,000 - (40 \times 2700)}{5} = 3400$

2.5 GRAPHICAL PRESENTATION OF DATA

It has already been discussed that one of the important functions of statistics is to present complex and unorganized (raw) data in such a manner that they would easily be understandable. According to King, 'One of the chief aims of statistical science is to render the meaning of masses of figures clear and comprehensible at a glance.' This is often best accomplished by presenting the data in a pictorial (or graphical) form.

The graphical (diagrammatical) presentation of data has many advantages. The following persons rightly observed that

- With but few exceptions, memory depends upon the faculty of our brains possess of forming visual images and it is this power of forming visual images which lies at the root of the utility of diagrammatic presentation. —R. L. A. Holmes
- Cold figures are uninspiring to most people. Diagrams help us to see the pattern and shape of any complex situation. Just as a map gives us a bird's eye-view of the wide stretch of a country, so diagrams help as visualise the whole meaning of the numerical complex at a single glance. —M. J. Moroney

Figures are not always interesting, and as their size and number increases they become uninteresting and confusing to such an extent that nobody would like to study them. The work of a statistician is to understand the data himself, and to put them in such a way that their importance may be known to every one. According to Calvin F. Schmid, 'Charts and graphs represent an extremely useful and flexible medium for explaining, interpreting and analysing numerical facts largely by means of points, lines, areas and other geometric forms and symbols. They make possible the presentation of quantitative data in a simple, clear, and effective manner and facilitate comparison of values, trends and relationships.'

2.5.1 Functions of a Graph

Graphic presentation of frequency distributions facilitate easy understanding of data presentation and interpretation. The shape of the graph offers easy answers to several questions. The same information can also be obtained from tabular presentation of a frequency distribution, but the same is not as effective in highlighting the essential characteristics as explicitly as is possible in the case of graphic presentation.

The shape of the graph gives an exact idea of the variations of the distribution trends. Graphic presentation, therefore, serves as an easy technique for quick and effective comparison between two or more frequency distributions. When the graph of one frequency distribution is superimposed on the other, the points of contrast regarding the type of distribution and the pattern of variation become quite obvious. All these advantages necessitate a clear understanding of the various forms of graphic representation of a frequency distribution.

2.5.2 Advantages and Limitations of Diagrams (Graphs)

According to P. Maslov, 'Diagrams are drawn for two purposes (i) to permit the investigator to graph the essence of the phenomenon he is observing, and (ii) to permit others to see the results at a glance, i.e. for the purpose of popularisation.'

Advantages Few of the advantages and usefulness of diagrams are as follows:

- (i) *Diagrams give an attractive and elegant presentation:* Diagrams have greater attraction and effective impression. People, in general, avoid figures, but are always impressed by diagrams. Since people see pictures carefully, their effect on the mind is more stable. Thus, diagrams give delight to the eye and add the spark of interest.
- (ii) *Diagrams leave good visual impact:* Diagrams have the merit of rendering any idea readily. The impression created by a diagram is likely to last longer in the minds of people than the effect created by figures. Thus diagrams have greater memorizing value than figures.
- (iii) *Diagrams facilitate comparison:* With the help of diagrams, comparisons of groups and series of figures can be made easily. While comparing absolute figures, the significance is not clear but when these are presented by diagrams, the comparison is easy. The technique of diagrammatic representation should not be used when comparison is either not possible or is not necessary.
- (iv) *Diagrams save time:* Diagrams present the set of data in such a way that their significance is known without loss of much time. Moreover, diagrams save time and effort which are otherwise needed in drawing inferences from a set of figures.

- (v) *Diagrams simplify complexity and depict the characteristics of the data:* Diagrams, besides being attractive and interesting, also highlight the characteristics of the data. Large data can easily be represented by diagrams and thus, without straining one's mind, the basic features of the data can be understood and inferences can be drawn in a very short time.

Limitations We often find tabular and graphical presentations of data in annual reports, newspapers, magazines, bulletins, and so on. But, inspite their usefulness, diagrams can also be misused. A few limitations of these as a tool for statistical analysis are as under:

- (i) They provide only an approximate picture of the data.
- (ii) They cannot be used as alternative to tabulation of data.
- (iii) They can be used only for comparative study.
- (iv) They are capable of representing only homogeneous and comparable data.

2.5.3 General Rules for Drawing Diagrams

To draw useful inferences from graphical presentation of data, it is important to understand how they are prepared and how they should be interpreted. When we say that 'one picture is worth a thousand words', it neither proves (nor disproves) a particular fact, nor is it suitable for further analysis of data. However, if diagrams are properly drawn, they highlight the different characteristics of data. The following general guidelines are taken into consideration while preparing diagrams:

Title: Each diagram should have a suitable title. It may be given either at the top of the diagram or below it. The title must convey the main theme which the diagram intends to portray.

Size: The size and portion of each component of a diagram should be such that all the relevant characteristics of the data are properly displayed and can be easily understood.

Proportion of length and breadth: An appropriate proportion between the length and breadth of the diagram should be maintained. As such there are no fixed rules about the ratio of length to width. However, a ratio of $\sqrt{2} : 1$ or 1.414 (long side) : 1 (short side) suggested by Lutz in his book *Graphic Presentation* may be adopted as a general rule.

Proper scale: There are again no fixed rules for selection of scale. The diagram should neither be too small nor too large. The scale for the diagram should be decided after taking into consideration the magnitude of data and the size of the paper on which it is to be drawn. The scale showing the values as far as possible, should be in even numbers or in multiples of 5, 10, 20, and so on. The scale should specify the size of the unit and the nature of data it represents, for example, 'millions of tonnes', in Rs thousand, and the like. The scale adopted should be indicated on both vertical and horizontal axes if different scales are used. Otherwise, it can be indicated at some suitable place on the graph paper.

Footnotes and source note: To clarify or elucidate any points which need further explanation but cannot be shown in the graph, footnotes are given at the bottom of the diagrams.

Index: A brief index explaining the different types of lines, shades, designs, or colours used in the construction of the diagram should be given to understand its contents.

Simplicity: Diagrams should be prepared in such a way that they can be understood easily. To keep it simple, too much information should not be loaded in a single diagram as it may create confusion. Thus if the data are large, then it is advisable to prepare more than one diagram, each depicting some identified characteristic of the same data.

2.6 TYPES OF DIAGRAMS

There are a variety of diagrams used to represent statistical data. Different types of diagrams, used to describe sets of data, are divided into the following categories:

- **Dimensional diagrams**
 - (i) One dimensional diagrams such as histograms, frequency polygons, and pie charts.
 - (ii) Two-dimensional diagrams such as rectangles, squares, or circles.
 - (iii) Three dimensional diagrams such as cylinders and cubes.
- **Pictograms or Ideographs**
- **Cartographs or Statistical maps**

2.6.1 One-Dimensional Diagrams

These diagrams are most useful, simple, and popular in the diagrammatic presentation of frequency distributions. These diagrams provides a useful and quick understanding of the *shape* of the distribution and its characteristics. According to Calvin F. Schmid, 'The simple bar chart with many variations is particularly appropriate for comparing the magnitude (or size) of coordinate items or of parts of a total. The basis of comparison in the bar is linear or one-dimensional.'

These diagrams are called one-dimensional diagrams because only the length (height) of the bar (not the width) is taken into consideration. Of course, width or thickness of the bar has no effect on the diagram, even then the thickness should not be too much otherwise the diagram would appear like a two-dimensional diagram.

Tips for Constructing a Diagram The following tips must be kept in mind while constructing one-dimensional diagrams:

- (i) The width of all the bars drawn should be same.
- (ii) The gap between one bar and another must be uniform.
- (iii) There should be a common base to all the bars.
- (iv) It is desirable to write the value of the variable represented by the bar at the top end so that the reader can understand the value without looking at the scale.
- (v) The frequency, relative frequency, or per cent frequency of each class interval is shown by drawing a rectangle whose base is the class interval on the horizontal axis and whose height is the corresponding frequency, relative frequency, or per cent frequency.
- (vi) The value of variables (or class boundaries in case of grouped data) under study are scaled along the horizontal axis, and the number of observations (frequencies, relative frequencies or percentage frequencies) are scaled along the vertical axis.

The one-dimensional diagrams (charts) used for graphical presentation of data sets are as follows:

- Histogram
- Frequency polygon
- Frequency curve
- Cumulative frequency distribution (Ogive)
- Pie diagram

Bar graph: A graphical device for depicting data that have been summarized in a frequency distribution, relative frequency distribution, or per cent frequency distribution.

Histograms (Bar Diagrams) These diagrams are used to graph both ungrouped and grouped data. In the case of an ungrouped data, values of the variable (the characteristic to be measured) are scaled along the horizontal axis and the number of observations (or frequencies) along the vertical axis of the graph. The plotted points are then connected by straight lines to enhance the shape of the distribution. The height of such boxes (rectangles) measures the number of observations in each of the classes.

Listed below are the various types of histograms:

- | | |
|-----------------------------------|-------------------------------------|
| (i) Simple bar charts | (v) Paired bar charts |
| (ii) Grouped (or multiple) charts | (vi) Sliding bar charts |
| (iii) Deviation bar charts | (vii) Relative frequency bar charts |
| (iv) Subdivided bar charts | (viii) Percentage bar charts |

For plotting a histogram of a grouped frequency distribution, the end points of class intervals are specified on the horizontal axis and the number of observations (or frequencies) are specified on the vertical axis of the graph. Often class mid-points are posted on the horizontal axis rather than the end points of class intervals. In either case, the width of each bar indicates the class interval while the height indicates the frequency of observations in that class. Figure 2.2 is a histogram for the frequency distribution given in Table 2.12 of Example 2.1.

Remarks: Bar diagrams are not suitable to represent long period time series.

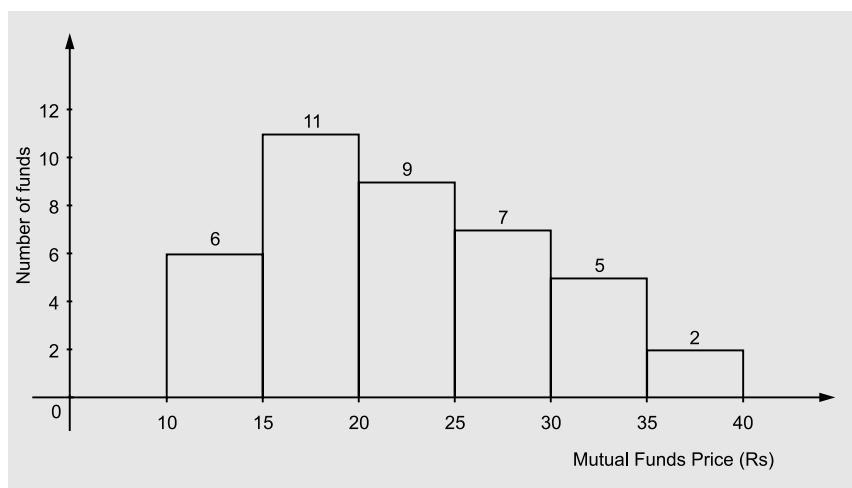


Figure 2.2
Histogram for Mutual Funds

Simple Bar Charts The graphic techniques described earlier are used for group frequency distributions. The graphic techniques presented in this section can also be used for displaying values of categorical variables. Such data is first tallied into summary tables and then graphically displayed as either *bar charts* or *pie charts*.

Bar charts are used to represent only one characteristic of data and there will be as many bars as number of observations. For example, the data obtained on the production of oil seeds in a particular year can be represented by such bars. Each bar would represent the yield of a particular oil seed in that year. Since the bars are of the same width and only the length varies, the relationship among them can be easily established.

Sometimes only lines are drawn for comparison of given variable values. Such lines are not thick and their number is sufficiently large. The different measurements to be shown should not have too much difference, so that the lines may not show too much dissimilarity in their heights.

Such charts are used to economize space, specially when observations are large. The lines may be either vertical or horizontal depending upon the type of variable—numerical or categorical.

Example 2.16: The data on the production of oil seeds in a particular year is presented in Table 2.35.

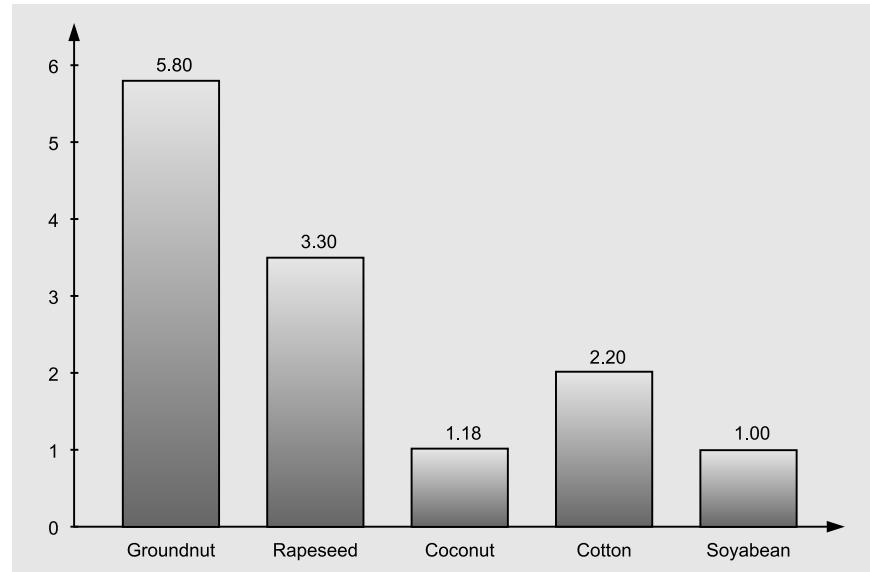
Table 2.35

Oil Seed	Yield (Million tonnes)	Percentage Production (Million tonnes)
Ground nut	5.80	43.03
Rapeseed	3.30	24.48
Coconut	1.18	8.75
Cotton	2.20	16.32
Soyabean	1.00	7.42
	13.48	100.00

Represent this data by a suitable bar chart.

Solution: The information provided in Table 2.35 is expressed graphically as the frequency bar chart as shown in Fig. 2.3. In this figure, each type of seed is depicted by a bar, the length of which represents the frequency (or percentage) of observations falling into that category.

Figure 2.3
Bar Chart Pertaining to Production of Oil Seeds



Remark: The bars should be constructed vertically (as shown in Fig. 2.3) when categorized observations are the outcome of a numerical variable. But if observations are the outcome of a categorical variable, then the bars should be constructed horizontally.

Example 2.17: An advertising company kept an account of response letters received each day over a period of 50 days. The observations were:

0	2	1	1	1	2	0	0	1	0	1	0	0	1	0	1	1	0
2	0	0	2	0	1	0	1	0	1	0	3	1	0	1	0	1	0
2	5	1	2	0	0	0	0	5	0	1	1	2	0				

Construct a frequency table and draw a line chart (or diagram) to present the data.

Solution: The observations are tallied into the summary table as shown in Table 2.35.

Figure 2.4 depicts a frequency bar chart for the number of letters received during a period of 50 days presented in Table 2.36.

Table 2.36 Frequency Distribution of Letters Received

Number of Letters Received	Tally	Number of Days (Frequency)
0		23
1		17
2		7
3		2
4	—	0
5		1
		50

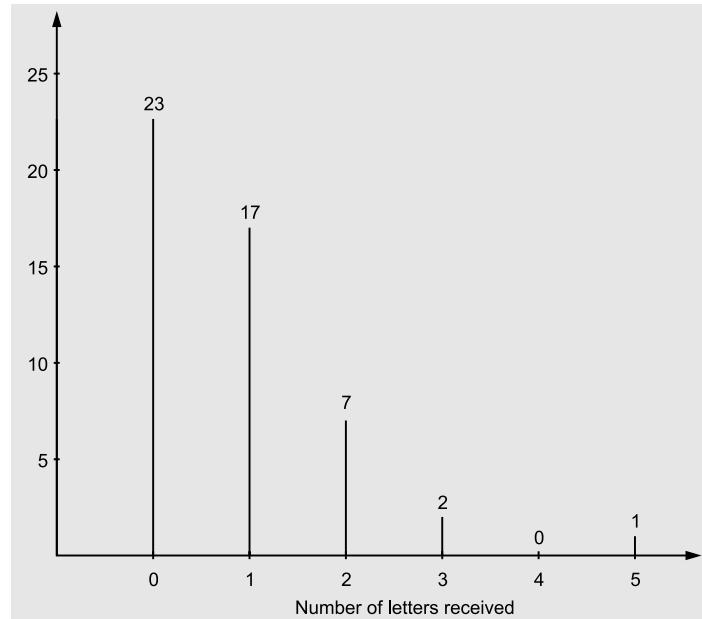


Figure 2.4
Number of Letters Received

Multiple Bar Charts A multiple bar chart is also known as grouped (or compound) bar chart. Such charts are useful for direct comparison between two or more sets of data. The technique of drawing such a chart is same as that of a single bar chart with a difference that each set of data is represented in different shades or colours on the same scale. An index explaining shades or colours must be given.

Example 2.18: The data on fund flow (in Rs crore) of an International Airport Authority during financial years 2001–02 to 2003–04 are given below:

	2001–02	2002–03	2003–04
Non-traffic revenue	40.00	50.75	70.25
Traffic revenue	70.25	80.75	110.00
Profit before tax	40.15	50.50	80.25

Represent this data by a suitable bar chart.

Solution: The multiple bar chart of the given data is shown in Fig. 2.5.

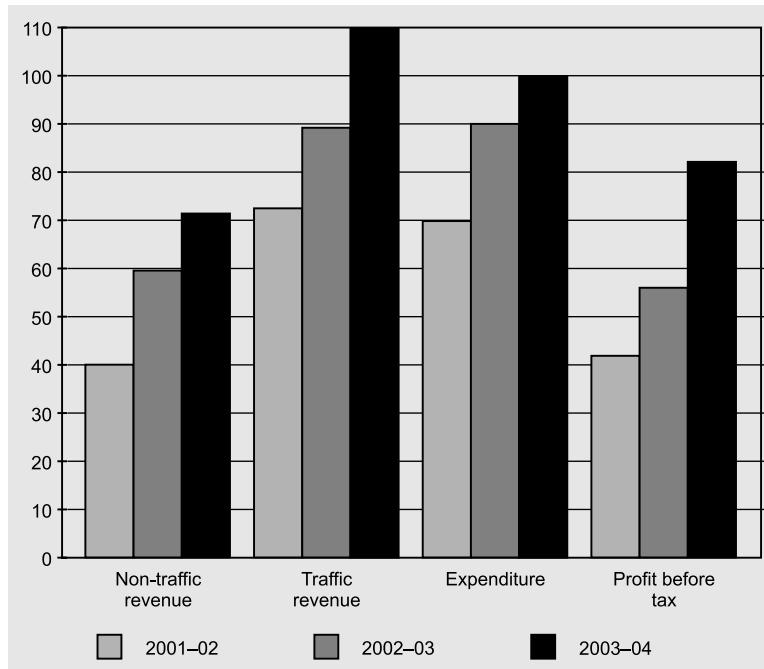


Figure 2.5
Multiple Bar Chart Pertaining to Performance of an International Airport Authority

Deviation Bar Charts Deviation bar charts are suitable for presentation of net quantities in excess or deficit such as profit, loss, import, or exports. The excess (or positive) values and deficit (or negative) values are shown above and below the base line.

Example 2.19: The following are the figures of sales and net profits of a company over the last three years.

(Per cent change over previous year)

Year	Sales Growth	Net Profit
2002–2003	15	30
2003–2004	12	53
2004–2005	18	-72

Present this data by a suitable bar chart.

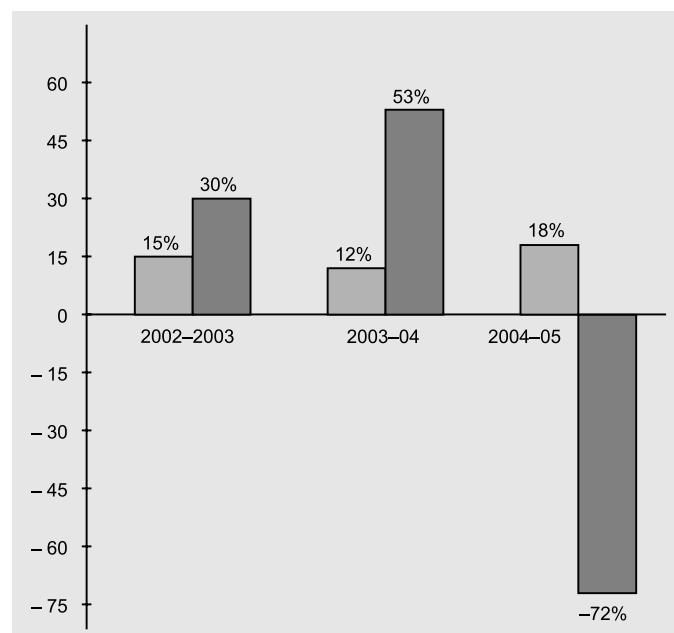
Solution: Figure 2.6 depicts deviation bar charts for sales and per cent change in sales over previous year's data.

Subdivided Bar Chart Subdivided bar charts are suitable for expressing information in terms of ratios or percentages. For example, net per capita availability of food grains, results of a college faculty-wise in last few years, and so on. While constructing these charts the various components in each bar should be in the same order to avoid confusion. Different shades must be used to represent various ratio values but the shade of each component should remain the same in all the other bars. An index of the shades should be given with the diagram.

A common arrangement while making these charts is that of presenting each bar in order of magnitude from the largest component at the base of the bar to the smallest at the end.

Since the different components of the bars do not start on the same scale, the individual bars are to be studied properly for their mutual comparisons.

Figure 2.6
Deviation Bar Chart Pertaining to
Sales and Profits



Example 2.20: The data on sales (Rs in million) of a company are given below:

	2002	2003	2004
Export	1.4	1.8	2.29
Home	1.6	2.7	2.9
Total	3.0	4.5	5.18

Solution: Figure 2.7 depicts a subdivided bar chart for the given data.

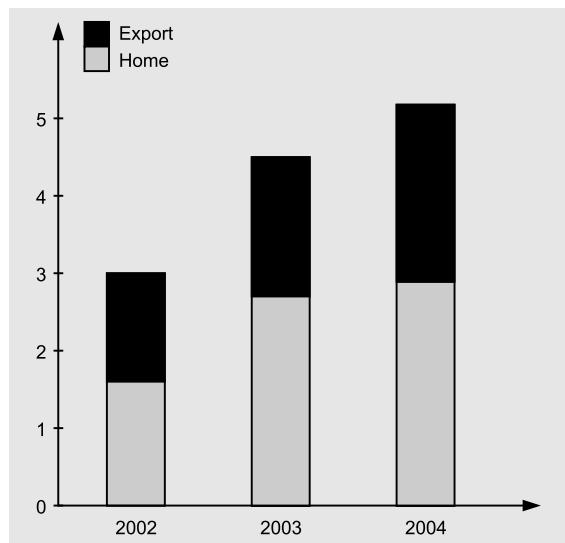


Figure 2.7
Subdivided Bar Chart Pertaining to Sales

Percentage Bar Charts When the relative proportions of components of a bar are more important than their absolute values, then each bar can be constructed with same size to represent 100%. The component values are then expressed in terms of percentage of the total to obtain the necessary length for each of these in the full length of the bars. The other rules regarding the shades, index, and thickness are the same as mentioned earlier.

Example 2.21: The following table shows the data on cost, profit, or loss per unit of a good produced by a company during the year 2003–04.

Particulars	2003			2004		
	Amount (Rs)	Percentage	Cumulative Percentage	Amount (Rs)	Percentage	Cumulative Percentage
Cost per unit						
(a) Labour	25	41.67	41.67	34	40.00	40.00
(b) Material	20	33.33	75.00	30	35.30	75.30
(c) Miscellaneous	15	25.00	100.00	21	24.70	100.00
Total cost	60	100		85	100	
Sales proceeds per unit	80	110		80	88	
Profit (+) or loss (-) per item	+ 20	+ 10		- 5	- 12	

Represent diagrammatically the data given above on percentage basis.

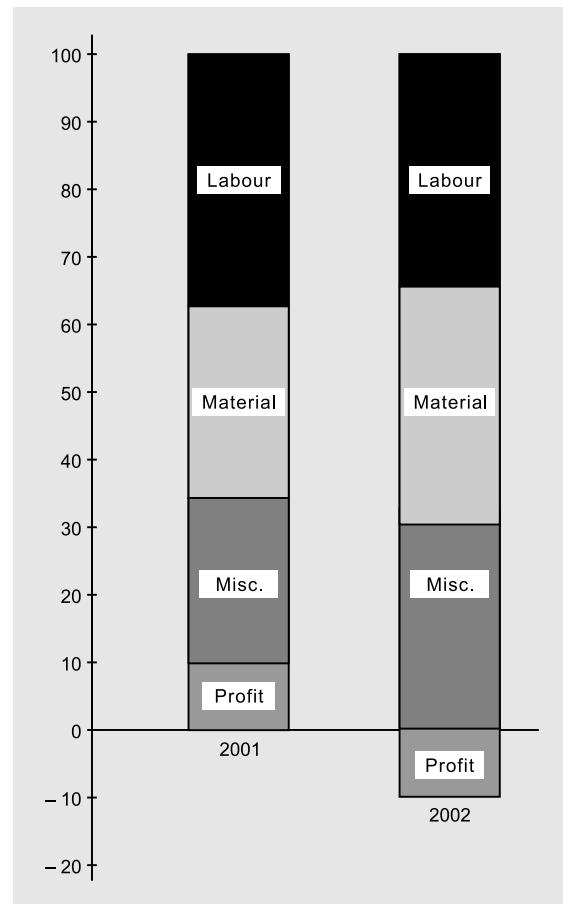
Solution: The cost, sales, and profit/loss data expressed in terms of percentages have been represented in the bar chart as shown in Fig. 2.8.

Frequency Polygon As shown in Fig. 2.9, the frequency polygon is formed by marking the mid-point at the top of horizontal bars and then joining these dots by a series of straight lines. The frequency polygons are formed as a closed figure with the horizontal axis, therefore a series of straight lines are drawn from the mid-point of the top base of the first and the last rectangles to the mid-point falling on the horizontal axis of the next outlaying interval with zero frequency. The frequency polygon is sometimes jagged in appearance.

A frequency polygon can also be converted back into a histogram by drawing vertical lines from the bounds of the classes shown on the horizontal axis, and then connecting them with horizontal lines at the heights of the polygon at each mid-point.

Figure 2.8

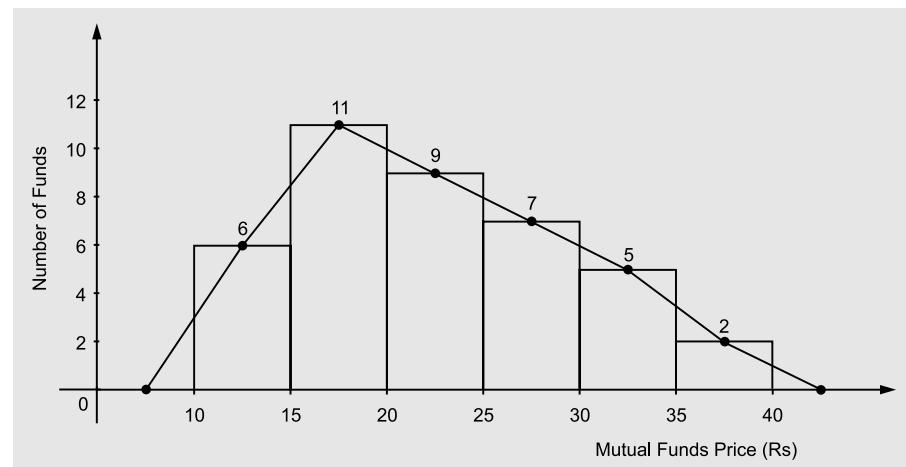
Percentage Bar Chart Pertaining to Cost, Sales, and Profit/Loss



Drawing a frequency polygon does not necessarily require constructing a histogram first. A frequency polygon can be obtained directly on plotting points above each class mid-point at heights equal to the corresponding class frequency. The points so drawn are then joined by a series of straight lines and the polygon is closed as explained earlier. In this case, horizontal x -axis measures the successive class mid-points and not the lower class limits. Figure 2.9 shows the frequency polygon for the frequency distribution presented by histogram in Fig. 2.2.

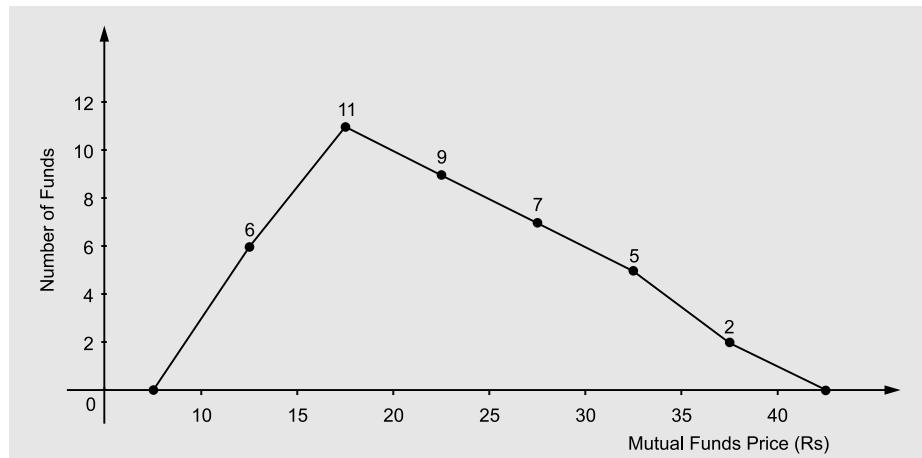
Figure 2.9

Frequency Polygon for Mutual Fund



Frequency Curve It is described as a smooth frequency polygon as shown in Fig. 2.10. A frequency curve is described in terms of its (i) symmetry (skewness) and its (ii) degree of peakedness (kurtosis).

Figure 2.10
Frequency Curve



Two frequency distributions can also be compared by superimposing two or more frequency curves provided the width of their class intervals and the total number of frequencies are equal for the given distributions. Even if the distributions to be compared differ in terms of total frequencies, they still can be compared by drawing per cent frequency curves where the vertical axis measures the per cent class frequencies and not the absolute frequencies.

Cumulative Frequency Distribution (Ogive) It enables us to see how many observations lie above or below certain values rather than merely recording the number of observations within intervals. Cumulative frequency distribution is another method of data presentation that helps in data analysis and interpretation. Table 2.37 shows the cumulative number of observations below and above the upper boundary of each class in the distribution.

A cumulative frequency curve popularly known as *Ogive* is another form of graphic presentation of a cumulative frequency distribution. The ogive for the cumulative frequency distribution given in Table 2.37 is presented in Fig. 2.11.

Once cumulative frequencies are obtained, the remaining procedure for drawing curve called ogive is as usual. The only difference being that the *y*-axis now has to be so scaled that it accommodates the total frequencies. The *x*-axis is labelled with the upper class limits in the case of less than ogive, and the lower class limits in case of more than ogive.

Table 2.37 Calculation of Cumulative Frequencies

Mutual Funds Price (Rs)	Upper Class Boundary	Number of Funds (f)	Cumulative Frequency	
			Less than	More than
10–15	15	6	6	40
15–20	20	11	6 + 11 = 17	40 – 6 = 34
20–25	25	9	17 + 9 = 26	34 – 11 = 23
25–30	30	7	26 + 7 = 33	23 – 9 = 14
30–35	35	5	33 + 5 = 38	14 – 7 = 7
35–40	40	2	38 + 2 = 40	7 – 5 = 2

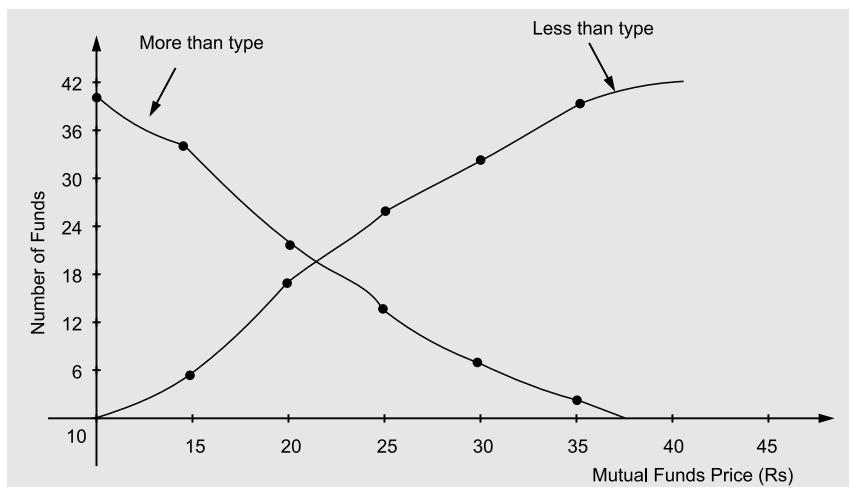


Figure 2.11
Ogive for Mutual Funds Prices

To draw a cumulative 'less than ogive', points are plotted against each successive upper class limit and a corresponding less than cumulative frequency value. These points are then joined by a series of straight lines and the resultant curve is closed at the bottom by extending it so as to meet the horizontal axis at the real lower limit of the first class interval.

To draw a cumulative 'more than ogive', points are plotted against each successive lower class limit and the corresponding more than cumulative frequency. These points are joined by a series of straight lines and the curve is closed at the bottom by extending it to meet the horizontal axis at the upper limit of the last class interval. Both the types of ogives so drawn are shown in Fig. 2.11.

It may be mentioned that a line drawn parallel to the vertical axis through the point of intersection of the two types of ogives will meet the x -axis at its middle point, and the value corresponding to this point will be the median of the distribution. Similarly, the perpendicular drawn from the point of intersection of the two curves on the vertical axis will divide the total frequencies into two equal parts.

Two ogives, whether *less than* or *more than* type, can be readily compared by drawing them on the same graph paper. The presence of unequal class intervals poses no problem in their comparison, as it does in the case of comparison of two frequency polygons. If the total frequencies are not the same in the two distributions, they can be first converted into per cent frequency distributions and then ogives drawn on a single graph paper to facilitate comparison.

Pie Diagram These diagrams are normally used to show the total number of observations of different types in the data set on a percentage basis rather than on an absolute basis through a circle. Usually the largest percentage portion of data in a pie diagram is shown first at 12 o'clock position on the circle, whereas the other observations (in per cent) are shown in clockwise succession in descending order of magnitude. The steps to draw a pie diagram are summarized below:

- (i) Convert the various observations (in per cent) in the data set into corresponding degrees in the circle by multiplying each by $3.6 (360 \div 100)$.
- (ii) Draw a circle of appropriate size with a compass.
- (iii) Draw points on the circle according to the size of each portion of the data with the help of a protractor and join each of these points to the center of the circle.

The pie chart has two distinct advantages: (i) it is aesthetically pleasing and (ii) it shows that the total for all categories or slices of the pie adds to 100%.

Example 2.22: The data shows market share (in per cent) by revenue of the following companies in a particular year:

Batata-BPL	30	Escorts-First Pacific	5
Hutchison-Essar	26	Reliance	3
Bharti-Sing Tel	19	RPG	2
Modi Dista Com	12	Srinivas	2
		Shyam	1

Draw a pie diagram for the above data.

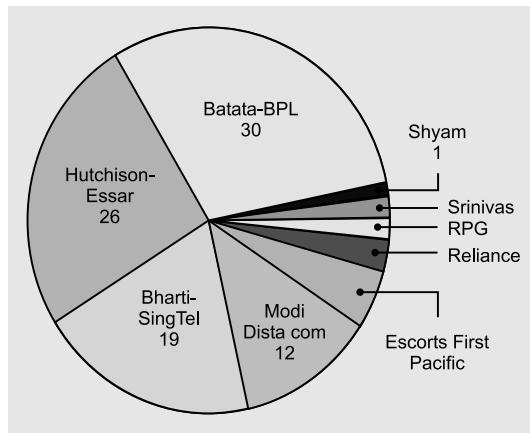
Solution: Converting percentage figures into angle outlay by multiplying each of them by 3.6 as shown in Table 2.38.

Table 2.38

Company	Market Share (Percent)	Angle Outlay (Degree)
Batata-BPL	30	108.0
Hutchison-Essar	26	93.6
Bharti-Sing Tel	19	68.4
Modi Dista Com	12	43.2
Escorts First Pacific	5	18.0
Reliance	3	10.8
RPG	2	7.2
Srinivas	2	7.2
Shyam	1	3.6
Total	100	360.0

Using the data given in Table 2.38 construct the pie chart displayed in Fig. 2.12 by dividing the circle into 9 parts according to degrees of angle at the centre.

Figure 2.12
Percentage Pie Chart



Example 2.23: The following data relate to area in millions of square kilometer of oceans of the world.

Ocean	Area (Million sq km)
Pacific	70.8
Atlantic	41.2
Indian	28.5
Antarctic	7.6
Arctic	4.8

Solution: Converting given areas into angle outlay as shown in Table 2.39.

Table 2.39

Ocean	Area (Million sq km)	Angle Outlay (Degrees)
Pacific	70.8	$\frac{70.8}{152.9} \times 360 = 166.70$
Atlantic	41.2	$\frac{41.2}{152.9} \times 360 = 97.00$
Indian	28.5	67.10
Antarctic	7.6	17.89
Arctic	4.8	11.31
Total	152.9	360.00

Pie diagram is shown in Fig. 2.13.

2.6.2 Two-Dimensional Diagrams

In one-dimensional diagrams or charts, only the length of the bar is taken into consideration. But in two-dimensional diagrams, both its height and width are taken into account for presenting the data. These diagrams, also known as *surface diagrams* or *area diagrams*, are:

- Rectangles
- Squares, and
- Circles.

Rectangles Since area of a rectangle is equal to the product of its length and width, therefore while making such type of diagrams both length and width are considered.

Rectangles are suitable for use in cases where two or more quantities are to be compared and each quantity is sub-divided into several components.

Example 2.24: The following data represent the income of two families A and B. Construct a rectangular diagram.

Item of Expenditure	Family A (Monthly Income Rs 30,000)		Family B (Monthly Income Rs 40,000)	
	Actual Expenses	Percentage of Expenses	Actual Expenses	Percentage of Expenses
Food	5550		7280	
Clothing	5100		6880	
House rent	4800		6480	
Fuel and light	4740		6320	
Education	4950		6640	
Miscellaneous	4860		6400	
Total	30,000		40,000	

Solution: Converting individual values into percentages taking total income as equal to 100 as shown in Table 2.40.

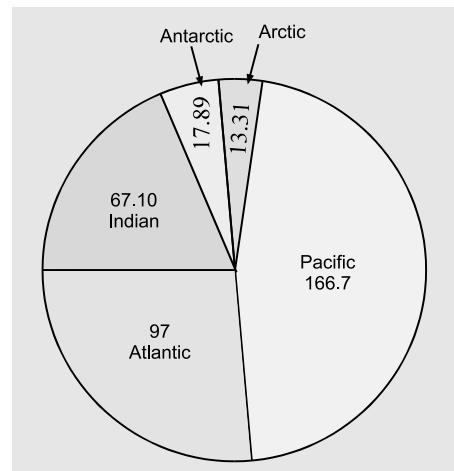
Table 2.40 Percentage Summary Table Pertaining to Expenses Incurred by Two Families

Item of Expenditure	Family A (Monthly Income Rs 3000)			Family B (Monthly Income Rs 4000)		
	Actual Expenses	Percentage of Expenses	Cumulative Percentage	Actual Expenses	Percentage of Expenses	Cumulative Percentage
Food	5550	18.50	18.50	7280	18.20	18.20
Clothing	5100	17.00	35.50	6880	17.20	35.40
House rent	4800	16.00	51.50	6480	16.20	51.60
Fuel and light	4740	15.80	6.78	6320	15.80	67.40
Education	4950	16.50	83.80	6640	16.60	84.00
Miscellaneous	4860	16.20	100.00	6400	16.00	100.00
Total	30,000	100.00		40,000	100.00	

The height of the rectangles shown in Fig. 2.14 is equal to 100. The difference in the total income is represented by the difference on the base line which is in the ratio 3 : 4.

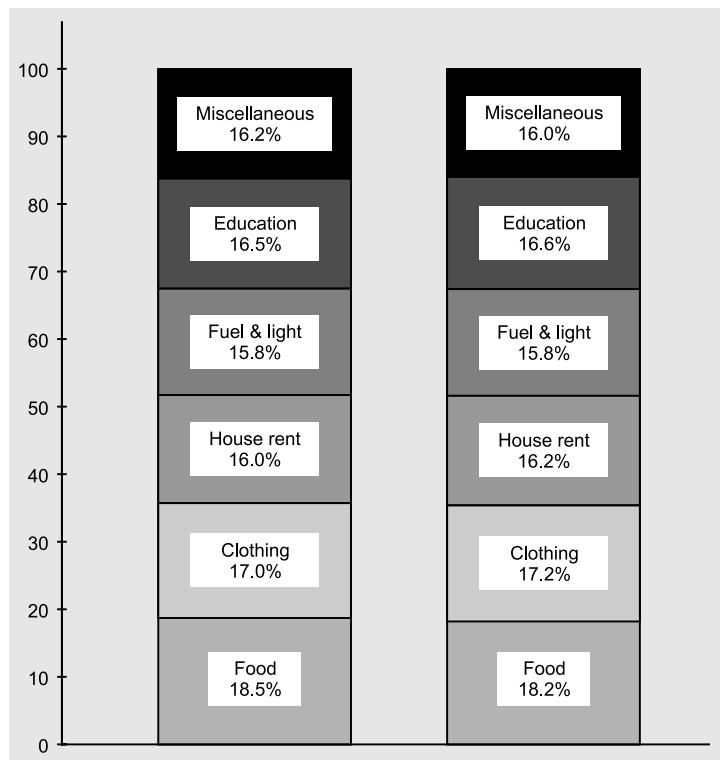
Squares Squares give a better comparison than rectangular bars when the difference of totals to be compared is large. For example, if in Example 2.24 the total expenses of families A and B are Rs 2000 and 20,000 respectively, then the width of the rectangles would be in the ratio 1 : 10. If such a ratio is taken, the diagram would look very unwieldy. Thus to overcome this difficulty, squares are constructed to make use of their areas to represent given data for comparison.

Figure 2.13
Per cent Pie Diagram



To construct a square diagram, first take the square-root of the values of various figures to be represented and then these values are divided either by the lowest figure or by some other common figure to obtain proportions of the sides of the squares. The squares constructed on these proportionate lengths must have either the base or the centre on a straight line. The scale is attached with the diagram to show the variable value represented by one square unit area of the squares.

Figure 2.14
Percentage of Expenditure by Two Families



Example 2.25: The following data represent the production (in million tonnes) of coal by different countries in a particular year.

Country	Production
USA	130.1
USSR	44.0
UK	16.4
India	3.3

Represent the data graphically by constructing a suitable diagram.

Solution: The given data can be represented graphically by square diagrams. For constructing the sides of the squares, the necessary calculations are shown in Table 2.41.

Table 2.41 Side of a Square Pertaining to Production of Coal

Country	Production (Million tonnes)	Square Root of Production Amount	Side of a Square (One square inch)
USA	130.1	11.406	1.267
USSR	44.0	6.633	0.737
UK	16.4	4.049	0.449
India	3.3	1.816	0.201

The squares representing the amount of coal production by various countries are shown in Fig. 2.15.

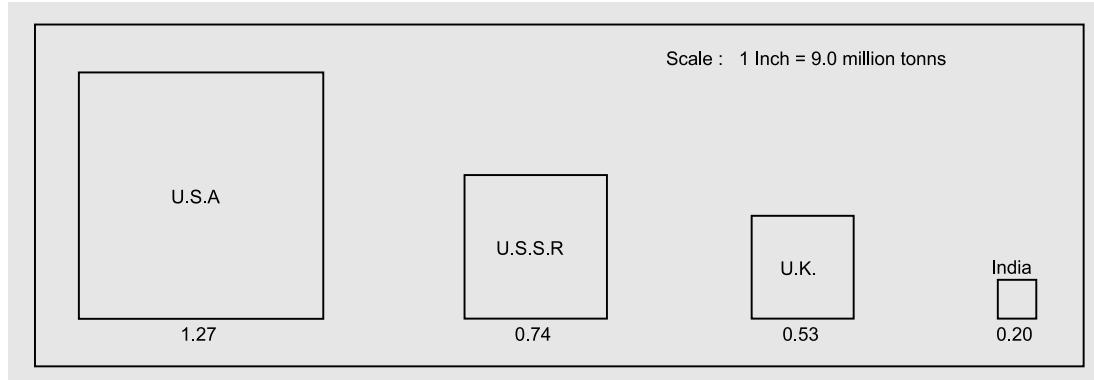


Figure 2.15
Coal Production in Different Countries

Circles Circles are alternatives, to squares to represent data graphically. The circles are also drawn such that their areas are in proportion to the figures represented by them. The circles are constructed in such a way that their centres lie on the same horizontal line and the distance between the circles are equal.

Since the area of a circle is directly proportional to the square of its radius, therefore the radii of the circles are obtain in proportion to the square root of the figures under representation. Thus, the lengths which were used as the sides of the square can also be used as the radii of circles.

Example 2.26: The following data represent the land area in different countries. Represent this data graphically using suitable diagram.

Country	Land Area (crore acres)
USSR	590.4
China	320.5
USA	190.5
India	81.3

Solution: The data can be represented graphically using circles. The calculations for constructing radii of circles are shown in Table 2.42.

Table 2.42 Radii of Circles Pertaining to Land Area of Countries

Country	Land Area (crore acres)	Square Root of Land Area	Radius of Circles (Inches)
USSR	590.4	24.3	0.81
China	320.5	17.9	0.60
USA	190.5	13.8	0.46
India	81.3	9.0	0.30

The various circles representing the land area of respective countries are shown in Fig. 2.16.

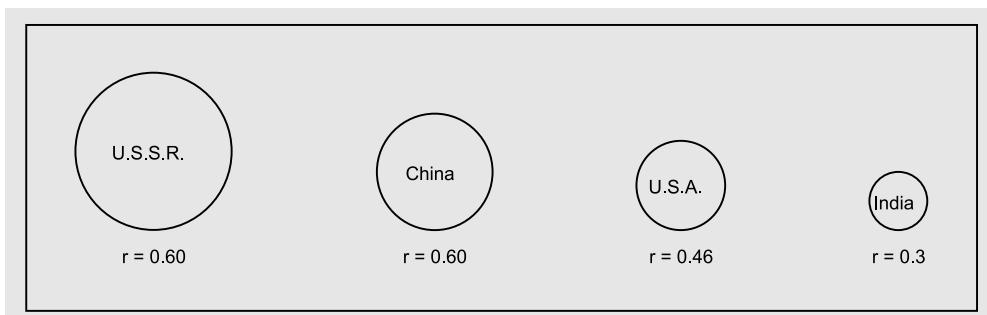


Figure 2.16
Land Area of Different Countries

2.6.3 Three-Dimensional Diagrams

Cylinders, spheres, cubes, and so on are known as three-dimensional diagrams because three dimensions—length, breadth, and depth, are taken into consideration for representing figures. These diagrams are used when only one point is to be compared and the ratio between the highest and the lowest measurements is more than 100 : 1. For constructing these diagrams, the cube root of various measurements is calculated and the side of each cube is taken in proportion to the cube roots.

Amongst the three-dimensional diagrams, cubes are the easiest and should be used only in those cases where the figures cannot be adequately presented through bar, square, or circle diagrams.

2.6.4 Pictograms or Ideographs

A pictogram is another form of pictorial bar chart. Such charts are useful in presenting data to people who cannot understand charts. Small symbols or simplified pictures are used to represent the size of the data. To construct pictograms or ideographs, the following suggestions are made:

- The symbols must be simple and clear.
- The quantity represented by a symbol should be given.
- Larger quantities are shown by increasing the number of symbols, and not by increasing the size of the symbols. A part of a symbol can be used to represent a quantity smaller than the whole symbol.

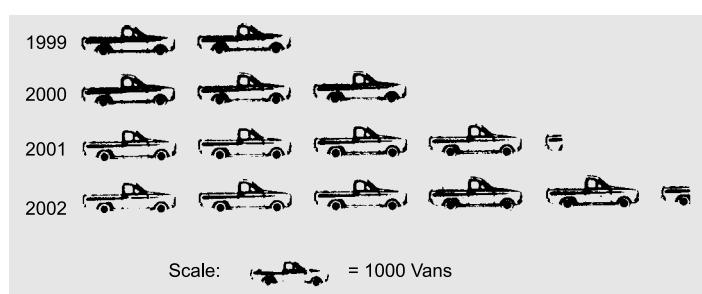
Example 2.27: Make a pictographic presentation of the output of vans during the year by a van manufacturing company.

Year	:	1999	2000	2001	2002
Output	:	2004	2996	4219	5324

Solution: Dividing the van output figures by 1000, we get 2.004, 2.996, 4.219, and 5.324 respectively.

Representing these figures by pictures of vans as shown in Fig. 2.17.

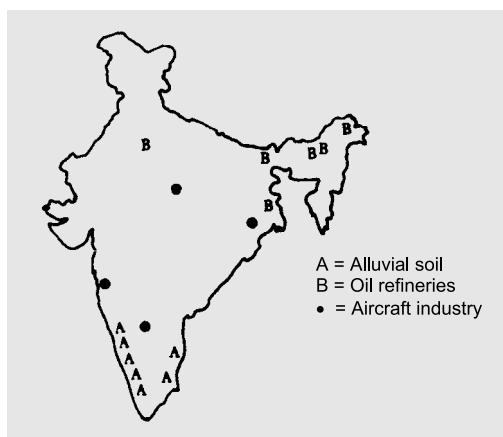
Figure 2.17
Output of Vans



2.6.5 Cartograms or Statistical Maps

Cartograms are used to represent graphical distribution of data on maps. The various figures in different regions on maps are shown either by (i) shades or colours, (ii) dots or bars, (iii) diagrams or pictures, or (iv) by putting numerical figures in each geographical area.

The following maps show the location of a particular type of soil, refineries, and aircraft industry in the country.



2.7 EXPLORATORY DATA ANALYSIS

This technique helps us to quickly describe and summarize a data set using simple arithmetic and diagrams. Such presentation of data values provides ways to determine relationships and trends, identify outliers and influential observations. In this section one of the useful techniques of exploratory data analysis, *stem-and-leaf displays* (or *diagrams*) technique is presented. This technique provides the *rank order* of the values in the data set and the shape of the distribution.

2.7.1. Stem-and-Leaf Displays

The stem-and-leaf display (or diagram) is another very simple but powerful technique to display quantitative data in a condensed form. The advantage of stem-and-leaf display over a frequency distribution is that the identity of each observation remains intact. This diagram provides us the rank order of the numerical values from lowest to highest in the data set and reveal the center, spread, shape and outliers (extremes) of a distribution. It is a graphical display of the numerical values in the data set and separates these values into *leading digits (or stem)* and *trailing digits (or leaves)*. The steps required to construct a stem-and-leaf diagram are as follows:

1. Divide each numerical value between the ones and the tens place. The number to the left is the stem and the number to the right is the leaf. The stem contains all but the last of the displayed digits of a numerical value. As with histogram, it is reasonable to have between 6 to 15 stems (each stem defines an interval of values). The stem should define equally spaced intervals. Stems are located along the vertical axis.

Sometimes numerical values in the data set are truncated or rounded off. For example, the number 15.69 is truncated to 15.6 but it is rounded off to 15.7.

2. List the stems in a column with a vertical line to their right.
3. For each numerical value, attach a leaf to the appropriate stem in the same row (horizontal axis). A leaf is the last of the displayed digits of a number. It is standard, but not mandatory, to put the leaves in increasing order at each stem value.
4. Provide a key to stem and leaf coding so that actual numerical value can be re-created, if necessary.

Remark If all the numerical values are three-digit integers, then to form a stem-and-leaf diagram, two approaches are followed:

- (i) Use the hundreds column as the stems and the tens column as the leaves and ignore the units column.
- (ii) Use the hundreds column as the stems and the tens column as the leaves after rounding of the units column.

Example 2.28 Consider the following marks obtained by 20 students in a business statistics test:

64	89	63	61	78	87	74	72	54	88
62	81	78	73	63	56	83	86	83	93

- (a) Construct a stem-and-leaf diagram for these marks to assess class performance
- (b) Describe the shape of this data set
- (c) Are there any outliers in this data set.

Solution (a) The numerical values in the given data set are ranging from 54 to 93. To construct a stem-and-leaf diagram, we make a vertical list of the stems (the first digit of each numerical value) as shown below:

Stem	Leaf
5	46
6	43123
7	84283
8	9781863
9	3

Rearrange all of the leaves in each row in rank order.

Stem	Leaf	9
5	46	9
6	12334	88
7	23488	7
8	1367889	6
9	3	6

4	8	8
3	8	7
3	4	6
6	2	3
6	2	3
4	1	2
4	1	2
1	3	

6 5 7 8 9

Each row in the diagram is a stem and numerical value on that stem is a leaf. For example, if we take the row 6/12334, it means there are five numerical values in the data set that begins with 6, i.e. 61, 62, 63, 63 and 64.

If the page is turned 90 degree clockwise and draw rectangles around the digits in each stem, we get a diagram similar to a histogram.

(b) Shape of the diagram is not symmetrical.

(c) There is no outlier (an observation far from the center of the distribution).

Example 2.29 The following data represent the annual family expenses (in thousand of rupees) on food items in a city.

13.8	14.1	14.7	15.2	12.8	15.6	14.9	16.7	19.2
14.9	14.9	14.9	15.2	15.9	15.2	14.8	14.8	19.1
14.6	18.0	14.9	14.2	14.1	15.3	15.5	18.0	17.2
17.2	14.1	14.5	18.0	14.4	14.2	14.6	14.2	14.8

Construct the stem-end-leaf diagram.

Solution: Since the annual costs (in Rs '000) in the data set all have two-digit integer numbers, the tens and units columns would be the leading digits and the remaining column (the tenth column) would be trailing digits as shown below:

Stem	Leaf	Stem	Leaf
12	8	12	8
13	8	13	8
14	17999988621142628	14	11122246678889999
15	2629235	15	2223569
16	7	16	7
17	2	17	2
18	000	18	000
19	21	19	12

Rearrange all the leaves in each row in the rank order as shown above.

Conceptual Questions 2C

22. What are the different types of charts known to you? What are their uses?
23. Point out the role of diagrammatic presentation of data. Explain briefly the different types of bar diagrams known to you.
24. Charts are more effective in attracting attention than other methods of presenting data. Do you agree? Give reasons for your answer. [MBA, HP Univ., 1998]
25. Discuss the utility and limitations (if any) of diagrammatic presentation of statistical data.
26. Diagrams are meant for a rapid view of the relation of different data and their comparisons. Discuss
27. Write short notes on pictographic and cartographic representations of statistical data.
28. What are the advantages of using a graph to describe a frequency distribution?
29. When constructing a graph of a grouped frequency distribution, is it necessary that the resulting distribution be symmetric? Explain.
30. Explain what is meant by a frequency polygon, a histogram, and a frequency curve.
31. Define the terms relative frequency and cumulative frequency. How are these related to a frequency distribution?
32. The distribution of heights of all students in the commerce department of the university has two peaks or is bimodal. The distribution of the IQs of the same students, however, has only one peak. How is this possible since the same students are considered in both cases? Explain.

Self-Practice Problems 2C

- 2.28 The following data represent the gross income, expenditure (in Rs lakh), and net profit (in Rs lakh)

during the years 1999 to 2002.

	1999–2000	2000–2001	2001–2002
Gross income	570	592	632
Gross expenditure	510	560	610
Net income	60	32	22

Construct a diagram or chart you prefer to use here.

- 2.29** Which of the charts would you prefer to represent the following data pertaining to the monthly income of two families and the expenditure incurred by them.

Expenditure on	Family A (Income Rs 17,000)	Family B (Income Rs 10,000)
Food	4000	5400
Clothing	2800	3600
House rent	3000	3500
Education	2300	2800
Miscellaneous	3000	5000
Saving or deficits	+1900	-300

- 2.30** The following data represent the outlays (Rs crore) by heads of development.

Heads of Development	Centre	States
Agriculture	4765	7039
Irrigation and Flood control	6635	11,395
Energy	9995	8293
Industry and Minerals	12,770	2985
Transport and Communication	12,200	5120
Social services	8216	1420
Total	54,581	36,252

Represent the data by a suitable diagram and write a report on the data bringing out the silent features.

- 2.31** Make a diagrammatic representation of the following textile production and imports.

	Value (in crore)	Length (in hundred yards)
Mill production	116.4	426.9
Handloom production	106.8	192.8
Imports	319.7	64.7

What conclusions do you draw from the diagram?

- 2.32** Make a diagrammatic representation of the following data:

Country	Production of Sugar in a Certain Year in Quintals (10,00,000)
Cuba	32
Australia	30
India	20
Japan	5
Java	1
Egypt	1

- 2.33** The following data represent the estimated gross area under different cereal crops during a particular year.

Crop	Gross Area ('000 hectares)	Crop	Gross Areas ('000 hectares)
Paddy	34,321	Ragi	2656
Wheat	18,287	Maize	6749
Jowar	22,381	Barley	4422
Bajra	15,859	Small millets	6258

Draw a suitable chart to represent the data.

- 2.34** The following data indicate the rupee sales (in '000) of three products according to region.

Product Group	Sales (in Rs '000)			Total Sales (Rs '000)
	North	South	East	
A	70	75	90	135
B	90	60	100	250
C	50	60	40	150
	210	195	230	533

- (i) Using vertical bars, construct a bar chart depicting total sales region-wise.
- (ii) Construct a component chart to illustrate the product breakdown of sales region-wise by horizontal bars.
- (iii) Construct a pie chart illustrating total sales.

- 2.35** The following data represent the income and dividend for the year 2000.

Year	Income Per Share (in Rs)	Dividend Per Share (in Rs)
1995	5.89	3.20
1996	6.49	3.60
1997	7.30	3.85
1998	7.75	3.95
1999	8.36	3.25
2000	9.00	4.45

- (i) Construct a line graph that indicates the income per share for the period 1995–2000.
- (ii) Construct a component bar chart that depicts dividends per share and retained earning per share for the period 1995–2000.
- (iii) Construct a percentage pie chart depicting the percentage of income paid as dividend. Also construct a similar percentage pie chart for the period 1998–2000. Observe any difference between the two pie charts.

- 2.36** The following time series data taken from the annual report of a company represents per-share net income, dividend, and retained earning during the period 1996–2000.

Source	1996	1997	1998	1999	2000
Net income (in Rs)	67.40	67.54	66.44	67.78	14.62
Dividends (in Rs)	8.08	8.28	8.40	8.50	8.75
Retained earnings (in Rs)	66.82	66.81	65.64	66.88(–)	10.56

- (i) Construct a bar chart for per-share income for the company during 1996–2000.

(ii) Construct a component bar chart depicting the allocation of annual earnings for the company during 1996–2000.

(iii) Construct a line graph for the per-share net income for the period 1996–2000.

2.37 The following data indicate the number of foreign tourists arrived in India during the period 1998–2001.

2.37 The following data indicate the number of foreign tourists arrived in India during the period 1998–2001.

Country	Number of Tourists Arrived		
	1998–1999	1999–2000	2000–2001
USA	2110	2340	2245
UK	5393	6245	6384
Middle East and Gulf	1114	1045	1097
Australia	2432	1849	2249
Western and Eastern Europe	5492	5890	5990

- (a) Construct a line graph for the arrival of tourists during 1998–2001.

(b) Construct a histogram for the data.

(c) Construct a percentage pie chart for the different countries.

2.38 Find a business or economic related data set of interest to you. The data set should be made up of at least 100 quantitative observations.

(a) Show the data in the form of a standard frequency distribution.

(b) Using the information obtained from part (i) briefly describe the appearance of your data.

2.39. The first row of a stem-end-leaf diagram appears as follows: 26/14489. Assume whole number values

(a) What is the possible range of values in this row?

Hints and Answers

- 2.39.** (a) 260 to 269 (b) 5
 (c) 261, 264, 264, 268, 269

2.40. (a)

9	147
10	02238
11	135566777
12	223489
13	02

(b) 91 94 97 100 102 102 103 108 111
 113 115 115 116 116 117 117 122 122
 123 124 128 129 130 132

- (b) How many data values are in this row?
(c) List the actual values in this row of data.

2.40. Given the following stem-end-leaf display representing the amount of CNG purchased in litres (with leaves in tenths litre) for a sample of 25 vehicles in Delhi.

9	714
10	82230
11	561776735
12	394282
13	20

- (a) Rearrange the leaves and form the revised stem-end-leaf display.
(b) Place the data into an ordered array.

2.41. The following stem-end-leaf display shows the number of units produced per day of in item an a factory.

3	8
4	-
5	6
6	0133559
7	0236778
8	59
9	00156
10	36

- (a) How many days were studied
 - (b) What are the smallest value and the largest value?
 - (c) List the actual values in second and fourth row.
 - (d) How many values are 80 or more.
 - (e) What is the middle value?

- 2.42.** A survey of the number of customer used PCO/STD both located at a college gate to make telephone calls last week revealed the following information

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

(a) Develop a stem-and-leaf display
(b) How many calls did a typical customer made?
(c) What were the largest and the smallest number of calls made?

- 2.41.** (a) 25 (b) 38, 106
 (c) No values, 60, 61, 63, 63, 65, 65, 69
 (d) 9 (e) 76 (f) 16

2.42. (a)	0	5
	1	28
	2	-
	3	0024789
	4	12366
	5	2

- (b) 16 customers were studied
 - (c) Number of customers visited ranged from 5 to 52.

Formulae Used

1. Class interval for a class in a frequency distribution

$$h = \text{Upper limit} - \text{Lower limit}$$
2. Midpoint of a class in a frequency distribution

$$m = \frac{\text{Upper limit} + \text{Lower limit}}{2}$$
3. Approximate interval size to be used in constructing a frequency distribution

$$h = \frac{\text{Largest data value} - \text{Smallest data value}}{\text{Number of class intervals}}$$

4. Approximate number of class intervals for constructing a frequency distribution: $2^k \geq N$, where k and N represent the number of classes and total number of observations, respectively.

Review Self-Practice Problems

- 2.39** If the price of a two-bed room flat in Gurgaon varies from Rs 9,00,000 to Rs 12,00,000, then
- (a) Indicate the class boundaries of 10 classes into which these values can be grouped
 - (b) What class interval width did you choose?
 - (c) What are the 10 class midpoints?
- 2.40** Students admitted to the MBA programme at FMS were asked to indicate their preferred major area of specialization. The following data were obtained

Area of Specialization	Number of Students
Marketing	50
Finance	22
HRD	08
Operations	10

- (a) Construct a relative and percentage frequency distribution.
(b) Construct a bar chart and pie chart.

- 2.41** The raw data displayed here are the scores (out of 100 marks) of a market survey regarding the acceptability of a new product launch by a company for a random sample of 50 respondents

40	45	41	45	45	30	39	8	48
25	26	9	23	24	26	29	8	40
41	42	39	35	18	25	35	40	42
43	44	36	27	32	28	27	25	26
38	37	36	35	32	28	40	41	43
44	45	40	39	41				

- (a) Form a frequency distribution having 9 class intervals
(b) Form a percentage distribution from the frequency distribution in part (a)
(c) Form a histogram of the frequency distribution in part (a)

- 2.42** State whether each of the following variables is qualitative or quantitative and indicate the measurement scale that is appropriate for each:

- | | |
|-----------------------|------------------------|
| (a) Age | (b) Gender |
| (c) Class rank | (d) Annual sales |
| (e) Method of payment | (f) Earnings per share |

- 2.43** The following data represent the sales of car tyres of various brands by a retail showroom of tyres during the year 2001–02.

Brand of Tyre	Tyres Sold
Dunlop	136
Modi	221
Firestone	138
Ceat	84
Goodyear	101
JK	120

- (a) Construct a bar chart and pie chart.
(b) Which of these charts do you prefer to use? Why?

- 2.44** The following data represent the expenditure incurred on following heads by a company during the year 2002

Expenditure Head	Amount (Rs in lakh)
Raw materials	1,689
Taxes	582
Manufacturing expenses	543
Employees salary	470
Depreciation	94
Dividend	75
Misc. expenses	286
Retained income	51

- (a) Construct a bar chart and pie chart.
(b) Which of these charts do you prefer to use? Why?

- 2.45** Draw an ogive by less than method and determine the number of companies earning profits between Rs 45 crore and Rs 75 crore:

Profit (Rs in crore)	Number of Companies
10–20	8
20–30	12
30–40	20
40–50	24
50–60	15
60–70	10
70–80	7
80–90	3
90–100	1

- 2.46** The following data represent the hottest career options in marketing:

Career Option	Percentage
Product Manager	23
Market Research Executive	10
Direct Marketing Manager	20
Manager-Events and Productions	10
VP Marketing	16
Other Marketing Careers	21

Develop the appropriate display(s) and thoroughly analyse the data.

- 2.47** Software engineers at a software development company want to adopt a flexitime system beginning at 7.00, 7.30, 8.00, and 9.00 a.m. The following data represent a sample of the starting times by the engineers.

7.00 8.00 7.30 8.30 8.30 9.00 8.30 7.00 7.30 8.30
7.30 7.30 8.00 8.00 9.00 8.00 8.30 8.30 7.00 8.30

Develop the appropriate display(s) of data and summarize your conclusions about preferences in the flexitime system.

- 2.48** The frequency distribution of GMAT scores from a sample of 50 applicants to an MBA course revealed that none of the applicants scored below 450 and that the table was formed by choosing class intervals $450 < 500$, $500 < 550$, and so on with the last class grouping $700 < 750$. If two applicants scored between $450 < 500$ and 16 applicants scored between $500 < 550$, then construct a percentage ogive and calculate the percentage of applicants scoring below 500, between 500 and 550, and below 750.

- 2.49** The data represent the closing prices of 40 common stocks.

29	34	43	8	37	8	7	30	35
19	9	16	38	53	16	1	48	18
9	9	10	37	18	8	28	24	21
18	33	31	32	29	79	11	38	11
52	14	9	33					

- (a) Construct frequency and relative frequency distributions for the data.
(b) Construct cumulative frequency and cumulative relative frequency distributions of the data.

- 2.50** An NGO working for environmental protection took water samples from twelve different places along the route of the Yamuna river from Delhi to Agra. These samples were tested in the laboratory and rated as to the amount of solid pollution suspended in each sample. The results of the testing are given below:

35.5	50.0	65.7	51.5	47.2	30.7	37.1
49.0	57.0	43.4	35.8	46.4		

- (a) Develop an appropriate display(s) of the data.
(b) If the pollution rating of 42 (ppm) indicates excessive pollution, then how many samples would be rated as having excessive pollution?

- 2.51** The noise level of aircraft departing from an airport was rounded to the nearest integer value and grouped in a frequency distribution having marks at 90 and 120 decibels.

The noise level below 90 decibels is not considered serious, while any level above 130 decibels is very serious and almost deafening. If the residents of the airport area are raising this issue and bringing it to the notice of the government, then is this distribution adequate for their concern?

- 2.52** The distribution of disability adjusted life year (DALY) loss by certain causes in 1990 (in percentage) is given below:

Cause	India	China	World
• Communicable diseases	50.00	25.30	45.80
• Non-communicable diseases	40.40	58.00	42.80
• Injuries	9.10	16.70	12.00

Depict this data by pie chart and bar chart.

[Delhi Univ., MBA, 1999]

- 2.53** A government hospital has the following data representing weight in kg at birth of 200 premature babies:

Weight	Number of Babies
0.5–0.7	14
0.8–1.0	16
1.1–1.3	25
1.4–1.6	26
1.7–1.9	28
2.0–2.2	36
2.3–2.5	37
2.6–2.8	18

- (a) Develop an appropriate display(s).
(b) Calculate the approximate middle value in the data set.
(c) If a baby below 2 kg is kept in the ICU as a precaution, then what percentage of premature babies need extra care in the ICU?

- 2.54** The medical superintendent of a hospital is concerned about the amount of waiting time for a patient before being treated in the OPD. The following data of waiting time (in minutes) were collected during a typical day:

Waiting Time	Number of Patients
30–40	125
40–50	195
50–60	305
60–70	185
70–80	120
80–90	70

- (a) Use the data to construct 'more than' and 'less than' frequency distributions and ogive.
- (b) Use the ogive to estimate how long 75 per cent of the patients should expect to wait.
- 2.55** A highway maintenance agency has ordered a study of the amount of time vehicles must wait at a toll gate of a recently constructed highway which is severely clogged and accident-prone in the morning. The following data were collected on the number of minutes that 950 vehicles waited in line a typical day.

Waiting Time	Number of Vehicles
1.00–1.39	30
1.40–1.79	42
1.80–2.19	75
2.20–2.59	110
2.60–2.99	120
3.00–3.39	130
3.40–3.79	108
3.80–4.19	94
4.20–4.59	85
4.60–4.99	78

Construct an ogive and determine what percentage of the vehicles had to wait more than three minutes in line?

Case Studies

Case 2.1: Housing Complex

The welfare committee of a large housing complex wants to understand the possibility of appointing private security guards at the entrance gate of the complex for 24-hour duty. There are 810 flats in the housing complex, and the owners were asked to vote for or against the proposal. The following data were collected:

<i>Should the guards be appointed</i>	
Yes	194
No	121
Not sure	73
No response	422

Questions for Discussion

- (a) Convert the data to percentages and construct (i) a bar chart and (ii) a pie chart. Which of these charts do you prefer to use? Why?
- (b) Eliminating the 'no response' group, convert the

remaining 388 responses to percentages and again construct bar and pie charts.

Suppose you have been designated as poll officer, based on your analysis of the data what would you like to suggest to the president of the welfare committee?

Case 2.2: Portfolio Management

A portfolio manager keeps a close watch on price-earnings ratios (defined as current market price divided by earnings for the most recent four quarters) of 200 common stocks. He reasons that, when the majority of stocks in a representative sample have low price-earnings (P-E) ratios by historical standards, it is time to become an aggressive buyer. Low P-Es may mean that investors in general are unrealistically pessimistic. Moreover stocks with low P-Es can benefit in a two-fold way when earnings increase: (a) higher earnings multiplied by a constant P-E ratio means a higher market price and (b) rising earnings are usually accompanied by rising P-E ratios.

Price-Earning Ratios for 200 Common Stocks							
11.1	12.6	26.7	5.2	8.3	5.5	6.8	7.6
7.3	18.1	14.6	10.9	7.2	9.5	9.2	11.8
12.0	16.9	10.1	14.6	5.2	7.5	11.1	19.9
14.9	7.4	6.0	39.9	29.3	35.1	6.8	39.0
6.1	6.2	26.8	33.7	9.6	16.6	10.9	11.2
22.6	46.0	7.3	29.7	10.3	6.4	9.6	7.6
10.3	5.0	14.4	11.6	8.3	7.9	17.8	7.5
7.8	7.3	8.0	20.2	5.6	8.3	7.7	10.7
8.6	14.5	6.0	5.4	12.6	14.8	9.2	14.1
15.7	10.4	7.0	11.0	6.3	8.4	7.6	16.9
7.9	8.3	13.1	9.8	8.2	18.0	26.6	7.8
4.1	10.6	15.3	7.2	35.5	6.1	10.2	6.1
7.8	8.1	30.0	15.0	6.1	15.4	10.1	9.6

6.8	4.4	6.8	9.1	16.3	5.4	5.9	6.5
7.9	44.9	13.8	12.3	10.9	9.3	11.9	10.0
7.6	17.9	7.1	8.4	35.5	7.4	7.7	8.3
15.8	8.3	23.1	8.4	12.4	7.8	8.2	9.8
13.7	15.8	4.7	7.9	26.4	6.2	11.4	13.2
8.6	11.7	8.6	13.7	9.3	16.6	8.7	39.7
14.0	9.1	7.1	10.9	23.4	13.3	10.9	24.0
11.9	8.7	15.6	27.7	10.4	16.9	6.9	5.5
22.8	8.5	22.2	5.8	14.7	8.0	7.5	10.5
4.4	7.1	63.8	12.5	13.3	10.5	5.5	16.0
53.1	7.4	24.1	15.3	29.1	11.0	9.9	36.3
9.6	6.6	5.1	7.8	8.4	38.3	20.4	9.1

Questions for Discussion

- (a) Organize the values of the variable into an array.
- (b) Construct a frequency distribution table.
- (c) Present the resulting frequency distribution as a histogram or frequency polygon and comment on the pattern.
- (d) Construct a cumulative frequency distribution and ogive.

This page is intentionally left blank.

We can easily represent things as we wish them to be...

—Aesop

If at first you do not succeed, you are just about average.

—Bill Cosby

Measures of Central Tendency

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- understand the role of descriptive statistics in summarization, description and interpretation of the data.
- understand the importance of summary measures to describe characteristics of a data set.
- use several numerical methods belonging to measures of central tendency to describe the characteristics of a data set.

3.1 INTRODUCTION

In Chapter 2, we discussed how raw data can be organized in terms of tables, charts, and frequency distributions in order to be easily understood and analysed. Although frequency distributions and corresponding graphical representations make raw data more meaningful, yet they fail to identify three major properties that describe a set of quantitative data. These three major properties are:

1. The numerical value of an observation (also called *central value*) around which most numerical values of other observations in the data set show a tendency to cluster or group, called the *central tendency*.
2. The extent to which numerical values are dispersed around the central value, called *variation*.
3. The extent of departure of numerical values from symmetrical (normal) distribution around the central value, called *skewness*.

These three properties—*central tendency*, *variation*, and *shape* of the frequency distribution—may be used to extract and summarize major features of the data set by the application of certain statistical methods called *descriptive measures* or *summary measures*. There are three types of summary measures:

1. Measures of central tendency
2. Measures of dispersion or variation
3. Measure of symmetry—skewness

Population parameter: A numerical value used as a summary measure using data of the population.

Sample statistic: A numerical value used as a summary measure using data of the sample for estimation or hypothesis testing.

These measures can also be used for comparing two or more populations in terms of the properties mentioned in the previous page to draw useful inferences.

The term ‘central tendency’ was coined because observations (numerical values) in most data sets show a distinct tendency to group or cluster around a value of an observation located somewhere in the middle of all observations. It is necessary to identify or calculate this typical *central value* (also called *average*) to describe or project the characteristic of the entire data set. This descriptive value is the measure of the *central tendency* or *location* and methods of computing this central value are called *measures of central tendency*.

If the descriptive summary measures are computed using data of samples, then these are called **sample statistic** or simply *statistic* but if these measures are computed using data of the population, they are called **population parameters** or simply *parameters*. The population parameter is represented by the Greek letter μ (read : mu) and sample statistic is represented by the Roman letter \bar{x} (read : x bar).

3.2 OBJECTIVES OF AVERAGING

A few of the objectives to calculate a typical central value or average in order to describe the entire data set are given below:

1. It is useful to extract and summarize the characteristics of the entire data set in a precise form. For example, it is difficult to understand individual families' need for water during summers. Therefore knowledge of the average quantity of water needed for the entire population will help the government in planning for water resources.
2. Since an ‘average’ represents the entire data set, it facilitates comparison between two or more data sets. Such comparison can be made either at a point of time or over a period of time. For example, average sales figures of any month can be compared with the preceding months, or even with the sales figures of competitive firms for the same months.
3. It offers a base for computing various other measures such as dispersion, skewness, kurtosis that help in many other phases of statistical analysis.

3.3 REQUISITES OF A MEASURE OF CENTRAL TENDENCY

The following are the few requirements to be satisfied by an average or a measure of central tendency:

1. **It should be rigidly defined** The definition of an average should be clear and rigid so that there must be uniformity in its interpretation by different decision-makers or investigators. There should not be any chance for applying discretion; rather it should be defined by an algebraic formula.
2. **It should be based on all the observations** To ensure that it should represent the entire data set, its value should be calculated by taking into consideration the entire data set.
3. **It should be easy to understand and calculate** The value of an average should be computed by using a simple method without reducing its accuracy and other advantages.
4. **It should have sampling stability** The value of an average calculated from various independent random samples of the same size from a given population should not vary much from another. The least amount of difference (if any) in the values is considered to be the sampling error.
5. **It should be capable of further algebraic treatment** The nature of the average should be such that it could be used for statistical analysis of the data set. For example, it should be possible to determine the average production in a particular year by the use of average production in each month of that year.

- 6. It should not be unduly affected by extreme observations** The value of an average should not be unduly affected by very small or very large observations in the given data. Otherwise the average value may not truly represent characteristics of the entire set of data.

3.4 MEASURES OF CENTRAL TENDENCY

The various measures of central tendency or averages commonly used can be broadly classified in the following categories:

1. Mathematical Averages

- (a) Arithmetic Mean commonly called the mean or average
 - Simple
 - Weighted
- (b) Geometric Mean
- (c) Harmonic Mean

2. Averages of Position

- (a) Median
- (b) Quartiles
- (c) Deciles
- (d) Percentiles
- (e) Mode

Notations

m_i = mid-point for the i th class in the data set

f_i = number of observations (or frequency) in the i th class; ($i = 1, 2, \dots, N$)

N = total number of observations in the population

n = number of observations in the sample (sample size)

l = lower limit of any class interval

h = width (or size) of the class interval

cf = cumulative frequency

Σ = summation (read: sigma) of all values of observations

3.5 MATHEMATICAL AVERAGES

Various methods of calculating mathematical averages of a data set are classified in accordance of the nature of data available, that is, ungrouped (unclassified or raw) or grouped (classified) data.

3.5.1 Arithmetic Mean of Ungrouped (or Raw) Data

There are two methods for calculating **arithmetic mean (A.M.)** for ungrouped or unclassified data:

- (i) Direct method, and
- (ii) Indirect or Short-cut method.

Direct Method

In this method A.M. is calculated by adding the values of all observations and dividing the total by the number of observations. Thus if x_1, x_2, \dots, x_N represent the values of N observations, then A.M. for a population of N observations is:

$$\text{Population mean, } \mu = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{1}{N} \sum_{i=1}^N x_i \quad (3-1a)$$

Mean: The sum of all the data values divided by their number.

However, for a sample containing n observations x_1, x_2, \dots, x_n , the sample A.M. can be written as:

$$\text{Sample mean, } \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3-1b)$$

The denominator of above two formulae is different because in statistical analysis the uppercase letter N is used to indicate the number of observations in the population, while the lower case letter n is used to indicate the number of observations in the sample.

Example 3.1: In a survey of 5 cement companies, the profit (in Rs lakh) earned during a year was 15, 20, 10, 35, and 32. Find the arithmetic mean of the profit earned.

Solution: Applying the formula (3-1b), we have

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{5} [15 + 20 + 10 + 35 + 32] = 22.4$$

Thus the arithmetic mean of the profit earned by these companies during a year was Rs 22.4 lakh.

Alternative Formula

In general, when observations x_i ($i = 1, 2, \dots, n$) are grouped as a frequency distribution, then A.M. formula (3-1b) should be modified as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n f_i x_i \quad (3-2)$$

where f_i represents the frequency (number of observations) with which variable x_i occurs in the given data set, i.e. $n = \sum_{i=1}^n f_i$.

Example 3.2: If A, B, C, and D are four chemicals costing Rs 15, Rs 12, Rs 8 and Rs 5 per 100 g, respectively, and are contained in a given compound in the ratio of 1, 2, 3, and 4 parts, respectively, then what should be the price of the resultant compound.

Solution: Using the formula (3-2), the sample arithmetic mean is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^4 f_i x_i = \frac{1 \times 15 + 2 \times 12 + 3 \times 8 + 4 \times 5}{1+2+3+4} = \text{Rs } 8.30$$

Thus the average price of the resultant compound should be Rs 8.30 per 100 g.

Example 3.3: The number of new orders received by a company over the last 25 working days were recorded as follows: 3, 0, 1, 4, 4, 4, 2, 5, 3, 6, 4, 5, 1, 4, 2, 3, 0, 2, 0, 5, 4, 2, 3, 3, 1. Calculate the arithmetic mean for the number of orders received over all similar working days.

Solution: Applying the formula (3-1b), the arithmetic mean is:

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^{25} x_i = \frac{1}{25} [3 + 0 + 1 + 4 + 4 + 4 + 2 + 5 + 3 + 6 + 4 \\ &\quad + 5 + 1 + 4 + 2 + 3 + 0 + 2 + 0 + 5 + 4 + 2 + 3 + 3 + 1] \\ &= \frac{1}{25} (71) = 2.84 \approx 3 \text{ orders (approx.)} \end{aligned}$$

Alternative approach: Use of formula (3-2)

Table 3.1 Calculations of Mean (\bar{x}) Value

Number of Orders (x_i)	Frequency (f_i)	$f_i x_i$
0	13	10
1	13	13
2	14	18
3	15	15
4	16	24
5	13	15
6	1	6
	25	71

$$\text{Arithmetic mean, } \bar{x} = \frac{1}{n} \sum f_i x_i = \frac{71}{25} = 2.8 \approx 3 \text{ orders (approx.)}$$

Example 3.4: From the following information on the number of defective components in 1000 boxes;

Number of defective components :	0	1	2	3	4	5	6
Number of boxes :	25	306	402	200	51	10	6

Calculate the arithmetic mean of defective components for the whole of the production line.

Solution: The calculations of mean defective components for the whole production line are shown in Table 3.2

Table 3.2 Calculations of \bar{x} for Ungrouped Data

Number of Defective Components (x_i)	Number of Boxes (f_i)	$f_i x_i$
0	25	0
1	306	306
2	402	804
3	200	600
4	51	204
5	10	50
6	6	36
	1000	2000

Applying the formula (3-2), the arithmetic mean is

$$\bar{x} = \frac{1}{n} \sum_{i=0}^6 f_i x_i = \frac{1}{1000} (2000) = 2 \text{ defective components.}$$

Short-Cut Method (Ungrouped Data)

In this method an arbitrary *assumed mean* is used as a basis for calculating deviations from individual values in the data set. Let A be the arbitrary assumed A.M. and let

$$d_i = x_i - A \quad \text{or} \quad x_i = A + d_i$$

Substituting the value of x_i in formula (3-1b), we have

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^n (A + d_i) = A + \frac{1}{n} \sum_{i=1}^n d_i \quad (3-3)$$

If frequencies of the numerical values are also taken into consideration, then the formula (3-3) becomes:

$$\bar{x} = A + \frac{1}{n} \sum_{i=1}^n f_i d_i \quad (3-4)$$

where $n = \sum_{i=1}^n f_i$ = total number of observations in the sample.

Example 3.5: The daily earnings (in rupees) of employees working on a daily basis in a firm are:

Daily earnings (Rs) :	100	120	140	160	180	200	220
Number of employees :	3	6	10	15	24	42	75

Calculate the average daily earning for all employees.

Solution: The calculations of average daily earning for employees are shown in Table 3.3.

Table 3.3 Calculations of \bar{x} for Ungrouped Data

Daily Earnings (in Rs) (x_i)	Number of Employees (f_i)	$d_i = x_i - A$ $= x_i - 160$	$f_i d_i$
100	3	- 60	- 180
120	6	- 40	- 240
140	10	- 20	- 200
(160) ← A	15	0	0
180	24	20	480
200	42	40	1680
220	75	60	4500
	175		6040

Here $A = 160$ is taken as assumed mean. The required A.M. using the formula (3-4) is given by

$$\bar{x} = A + \frac{1}{n} \sum_{i=1}^7 f_i d_i = 160 + \frac{6040}{175} = \text{Rs } 194.51$$

Example 3.6: The human resource manager at a city hospital began a study of the overtime hours of the registered nurses. Fifteen nurses were selected at random, and following overtime hours during a month were recorded:

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

Calculate the arithmetic mean of overtime hours during the month.

Solution: Calculations of arithmetic mean of overtime hours are shown in Table 3.4

Table 3.4 Calculations of \bar{x} for Ungrouped Data

Overtime Hours (x_i)	Number of Number (f_i)	$d_i = x_i - A$ $= x_i - 10$	$f_i d_i$
5	3	- 5	- 15
6	4	- 4	- 16
7	2	- 3	- 6
9	4	- 1	- 4
(10) ← A	2	0	0
12	5	2	10
13	3	3	9
15	2	5	10
	25		- 12

Here $A=10$ is taken as assumed mean. The required arithmetic mean of overtime using the formula (3-4) is as follows:

$$\bar{x} = A + \frac{1}{n} \sum_{i=1}^{25} f_i d_i = 10 - \frac{12}{25} = 9.52 \text{ hours}$$

3.5.2 Arithmetic Mean of Grouped (or Classified) Data

Arithmetic mean for grouped data can also be calculated by applying any of the following methods:

- (i) Direct method, and
- (ii) Indirect or Step-deviation method

For calculating arithmetic mean for a grouped data set, the following assumptions are made:

- (i) The class intervals must be closed
- (ii) The width of each class interval should be equal
- (iii) The values of the observations in each class interval must be uniformly distributed between its lower and upper limits.
- (iv) The mid-value of each class interval must represent the average of all values in that class, that is, it is assumed that all values of observations are evenly distributed between the lower and upper class limits.

Direct Method

The formula used in this method is same as formula (3-2) except that x_i is replaced with the mid-point value m_i of class intervals. The new formula becomes:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n f_i m_i \quad (3-5)$$

where m_i = mid-value of i th class interval.

f_i = frequency of i th class interval.

n = $\sum f_i$, sum of all frequencies

Mean value: A measure of central location (tendency) for a data set such that the observations in the data set tend to cluster around it.

Example 3.7: A company is planning to improve plant safety. For this, accident data for the last 50 weeks was compiled. These data are grouped into the frequency distribution as shown below. Calculate the A.M. of the number of accidents per week.

Number of accidents :	0–4	5–9	10–14	15–19	20–24
Number of weeks :	5	22	13	8	2

Solution: The calculations of A.M. are shown in Table 3.5 using formula (3-5).

Table 3.5 Arithmetic Mean of Accidents

Number of Accidents	Mid-value (m_i)	Number of Weeks (f_i)	$f_i m_i$
0–4	2	5	10
5–9	7	22	154
10–14	12	13	156
15–19	17	8	136
20–24	22	2	44
		50	500

The A.M. of the number of accidents per week is:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^5 f_i m_i = \frac{500}{50} = 10 \text{ accidents per week.}$$

Step-deviation Method

The formula (3-5) for calculating A.M. can be improved as formula (3-6). This improved formula is also known as the *step-deviation method*:

$$\bar{x} = A + \frac{\sum f_i d_i}{n} \times h \quad (3-6)$$

where A = assumed value for the A.M.

n = $\sum f_i$, sum of all frequencies

h = width of the class intervals

m_i = mid-value of i th class-interval

$$d_i = \frac{m_i - A}{h}, \text{ deviation from the assumed mean}$$

The formula (3-6) is very useful in those cases where mid-values (m_i) and/or frequencies (f_i) are in three or more digits. The calculation of d_i from m_i involves reducing each m_i by an amount A (called assumed arithmetic mean) and then dividing the reduced values by h (width of class intervals). This procedure is usually referred to as *change of location and scale or coding*.

Example 3.8: Calculate the arithmetic mean of accidents per week by the short cut method using the data of Example 3.7.

Solution: The calculations of the average number of accidents are shown in the Table 3.6.

Table 3.6 Arithmetic Mean of Accidents

Number of Accidents	Mid-value (m_i)	$d_i = (m_i - A)/h$ $= (m_i - 12)/5$	Number of Weeks (f_i)	$f_i d_i$
0–14	2	-2	5	-10
5–19	7	-1	22	-22
10–14	12	0	13	0
15–19	17	1	8	8
20–24	22	2	2	4
			50	-20

$$\begin{aligned} \text{The arithmetic mean } \bar{x} &= A + \left\{ \frac{1}{n} \sum f_i d_i \right\} h \\ &= 12 + \left\{ \frac{1}{50} (-20) \right\} 5 = 10 \text{ accidents per week} \end{aligned}$$

Example 3.9: The following distribution gives the pattern of overtime work done by 100 employees of a company. Calculate the average overtime work done per employee.

Overtime hours :	10–15	15–20	20–25	25–30	30–35	35–40
Number of employees :	11	20	35	20	8	6

Solution: The calculations of the average overtime work done per employee with assumed mean, $A = 22.5$ and $h = 5$ are given in Table 3.7.

Table 3.7 Calculations of Average Overtime

Overtime (hrs) x_i	Number of Employees, f_i	Mid-value (m_i)	$d_i = (m_i - 22.5)/5$	$f_i d_i$
10–15	11	12.5	-2	-22
15–20	20	17.5	-1	-20
20–25	35	22.5	0	0
25–30	20	27.5	1	20
30–35	8	32.5	2	16
35–40	6	37.5	3	18
	100			12

$$\text{The required A.M. is, } \bar{x} = A + \frac{\sum f_i d_i}{n} \times h = 22.5 + \frac{12}{100} \times 5 = 23.1 \text{ hrs}$$

Example 3.10: The following is the age distribution of 1000 persons working in an organization

Age Group	Number of Persons	Age Group	Number of Persons
20–25	30	45–50	105
25–30	160	50–55	70
30–35	210	55–60	60
35–40	180	60–65	40
40–45	145		

Due to continuous losses, it is desired to bring down the manpower strength to 30 per cent of the present number according to the following scheme:

- Retrench the first 15 per cent from the lower age group.
- Absorb the next 45 per cent in other branches.
- Make 10 per cent from the highest age group retire permanently, if necessary.

Calculate the age limits of the persons retained and those to be transferred to other departments. Also find the average age of those retained. [Delhi Univ., MBA; 2003]

Solution: (a) The first 15 per cent persons to be retrenched from the lower age groups are $(15/100) \times 1000 = 150$. But the lowest age group 20–25 has only 30 persons and therefore the remaining, $150 - 30 = 120$ will be taken from next higher age group, that is, 25–30, which has 160 persons.

(b) The next 45 per cent, that is, $(45/100) \times 1000 = 450$ persons who are to be absorbed in other branches, belong to the following age groups:

Age Groups	Number of Persons
25–30	$(160 - 120) = 40$
30–35	210
35–40	180
40–45	$(450 - 40 - 210 - 180) = 20$

(c) Those who are likely to be retired are 10 per cent, that is, $(10/100) \times 1000 = 100$ persons and belong to the following highest age groups:

Age Group	Number of Persons
55–60	$(100 - 40) = 60$
60–65	40

Hence, the calculations of the average age of those retained and/or to be transferred to other departments are shown in Table 3.8:

Table 3.8 Calculations of Average Age

Age Group (x_i)	Mid value, (m_i)	Number of Persons (f_i)	$d_i = (x_i - 47.5)/5$	$f_i d_i$
40–45	42.5	$145 - 20 = 125$	-1	-125
45–50	47.5 ← A	105	0	0
50–55	52.5	70	1	70
		300		-55

The required average age is, $\bar{x} = A + \frac{\sum d_i f_i}{n} \times h = 47.5 - \frac{55}{300} \times 5 = 46.58 = 47$ years (approx.).

3.5.3 Some Special Types of Problems and Their Solutions

Case 1: Frequencies are Given in Cumulative Form, that is, either ‘More Than Type’ or ‘Less Than Type’

As we know that the ‘more than type’ cumulative frequencies are calculated by adding frequencies from bottom to top, so that the first class interval has the highest cumulative frequency and it goes on decreasing in subsequent classes. But in case of ‘less than cumulative frequencies’, the cumulation is done downward so that the first class interval has the lowest cumulative frequency and it goes on increasing in the subsequent classes.

In both of these cases, data are first converted into inclusive class intervals or exclusive class intervals. Then the calculations for \bar{x} are done in the usual manner as discussed earlier.

Example 3.11: Following is the cumulative frequency distribution of the preferred length of kitchen slabs obtained from the preference study on housewives:

Length (in metres) more than :	1.0	1.5	2.0	2.5	3.0	3.5
Preference of housewives :	50	48	42	40	10	5

A manufacturer has to take a decision on what length of slabs to manufacture. What length would you recommend and why?

Solution: The given data are converted into exclusive class intervals as shown in Table 3.9. The frequency of each class has been found out by deducting the given cumulative frequency from the cumulative frequency of the previous class:

Table 3.9 Conversion into Exclusive Class Intervals

Length (in metres)	Preference of Housewives more than	Class Interval	Frequency
1.0	50	1.0–1.5	(50 – 48) = 2
1.5	48	1.5–2.0	(48 – 42) = 6
2.0	42	2.0–2.5	(42 – 40) = 2
2.5	40	2.5–3.0	(40 – 10) = 30
3.0	10	3.0–3.5	(10 – 5) = 5
3.5	5		

The calculations for mean length of slab are shown in Table 3.10.

Table 3.10 Calculations of Mean Length of Slab

Class Interval	Mid-value (m_i)	Preference of Housewives (f_i)	$d_i = \frac{m_i - 2.25}{0.5}$	$f_i d_i$
1.0–1.5	1.25	2	-2	-4
1.5–2.0	1.75	6	-1	-6
2.0–2.5	2.25	2	0	0
2.5–3.0	2.75	30	1	30
3.0–3.5	3.25	5	2	10
		45		30

$$\text{The mean length of the slab is } \bar{x} = A + \frac{\sum f_i d_i}{n} \times h = 2.25 + \frac{30}{45} \times 0.5 = 2.58$$

metres

Example 3.12: In an examination of 675 candidates, the examiner supplied the following information:

Marks Obtained (Percentage)	Number of Candidates	Marks Obtained (Percentage)	Number of Candidates
Less than 10	7	Less than 50	381
Less than 20	39	Less than 60	545
Less than 30	95	Less than 70	631
Less than 40	201	Less than 80	675

Calculate the mean percentage of marks obtained.

Solution: Arranging the given data into inclusive class intervals as shown in Table 3.11:

Table 3.11 Calculations of Mean Percentage of Marks

Marks Obtained (Percentage)	Cumulative Frequency	Class-intervals	Frequency
Less than 10	7	0–10	7
Less than 20	39	10–20	$(39 - 7) = 32$
Less than 30	95	20–30	$(95 - 39) = 56$
Less than 40	201	30–40	$(201 - 95) = 106$
Less than 50	381	40–50	$(381 - 201) = 180$
Less than 60	545	50–60	$(545 - 381) = 164$
Less than 70	631	60–70	$(631 - 545) = 86$
Less than 80	675	70–80	$(675 - 631) = 44$

The calculations for mean percentage of marks obtained by the candidates are shown in Table 3.12.

Table 3.12 Calculations of Mean Percentage of Marks

Class Intervals	Mid-value (m_i)	Number of Candidates (f_i)	$d_i = \frac{m_i - 35}{10}$	$f_i d_i$
0–10	5	7	-3	-21
10–20	15	32	-2	-64
20–30	25	56	-1	-56
30–40	(35) ← A	106	0	0
40–50	45	180	1	180
50–60	55	164	2	328
60–70	65	86	3	258
70–80	75	44	4	176
		675		801

The mean percentage of marks obtained is:

$$\bar{x} = A + \frac{\sum f_i d_i}{n} \times h = 35 + \frac{801}{675} \times 10 = 46.86 \text{ marks}$$

Case 2: Frequencies are not Given but have to be Calculated From the Given Data

Example 3.13: 168 handloom factories have the following distribution of average number of workers in various income groups:

Income groups	:	800–1000	1000–1200	1200–1400	1400–1600	1600–1800
Number of firms	:	40	32	26	28	42
Average number of workers	:	8	12	8	8	4

Find the mean salary paid to the workers.

Solution: Since the total number of workers (i.e. frequencies) working in different income groups are not given, therefore these have to be determined as shown in Table 3.13:

Table 3.13

Income Group (x_i)	Mid-values (m_i)	$d_i = \frac{m_i - A}{h}$ = $\frac{m_i - 1300}{200}$	Number of Firms (3)	Average Number of Workers (4)	Frequencies (f_i) (5) = (3) \times (4)	$m_i f_i$
(1)	(2)		(3)	(4)		
800–1000	900	-2	40	8	320	-640
1000–1200	1100	-1	32	12	384	-384
1200–1400	1300 ← A	0	26	8	208	0
1400–1600	1500	1	28	8	224	224
1600–1800	1700	2	42	4	168	336
			168	40	1304	-464

The required A.M. is given by

$$\bar{x} = A + \frac{\sum m_i f_i}{n} \times h = 1300 - \frac{464}{1304} \times 200 = 1228.84$$

Example 3.14: Find the missing frequencies in the following frequency distribution. The A.M. of the given data is 11.09.

Class Interval	Frequency	Class	Frequency
9.3–9.7	2	11.3–11.7	14
9.8–10.2	5	11.8–12.2	6
10.3–10.7	f_3	12.3–12.7	3
10.8–11.2	f_4	12.8–13.2	1
			60

Solution: The calculations for A.M. are shown in Table 3.14.

Table 3.14

Class Interval	Frequency (f_i)	Mid-value (m_i)	$d_i = \frac{m_i - 11.0}{0.5}$	$f_i d_i$
9.3–9.7	2	9.5	-3	-6
9.8–10.2	5	10.0	-2	-10
10.3–10.7	f_3	10.5	-1	$-f_3$
10.8–11.2	f_4	11.0 ← A	0	0
11.3–11.7	14	11.5	1	14
11.8–12.2	6	12.0	2	12
12.3–12.7	3	12.5	3	9
12.8–13.2	1	13.0	4	4
	60			$23 - f_3$

where the assumed mean is, $A = 11$. Applying the formula

$$\bar{x} = A + \frac{\sum f_i d_i}{n} \times h$$

we get $11.09 = 11.0 + \frac{23 - f_3}{60} \times 0.5$

or $0.09 = \frac{23 - f_3}{120}$ or $f_3 = 23 - 0.09 \times 120 = 12.2$

Since the total of the frequencies is 60 and $f_3 = 12.2$, therefore

$$f_4 = 60 - (2 + 5 + 12.2 + 14 + 6 + 3 + 1) = 16.8$$

Case 3: Complete Data are Not Given

Example 3.15: The pass result of 50 students who took a class test is given below:

Marks	:	40	50	60	70	80	90
Number of students	:	8	10	9	6	4	3

If the mean marks for all the students was 51.6, find out the mean marks of the students who failed.

Solution: The marks obtained by 40 students who passed are given in Table 3.15

Table 3.15

Marks	Frequency (f_i)	$f_i x_i$
40	8	320
50	10	500
60	9	540
70	6	420
80	4	320
90	3	270
	40	2370

$$\text{Total marks of all the students} = 50 \times 51.6 = 2580$$

$$\text{Total marks of 40 students who passed} = \sum f_i x_i = 2370$$

$$\text{Thus, marks of the remaining 10 students} = 2580 - 2370 = 210$$

$$\text{Hence, the average marks of 10 students who failed are } 210/10 = 21 \text{ marks}$$

Case 4: Incorrect Values have been used for the Calculation of Arithmetic Mean

Example 3.16: (a) The average dividend declared by a group of 10 chemical companies was 18 per cent. Later on, it was discovered that one correct figure, 12, was misread as 22. Find the correct average dividend.

(b) The mean of 200 observations was 50. Later on, it was found that two observations were misread as 92 and 8 instead of 192 and 88. Find the correct mean.

Solution: (a) Given $n = 10$ and $\bar{x} = 18$ per cent. We know that

$$\bar{x} = \frac{\sum x}{n} \quad \text{or} \quad \sum x = n \bar{x} = 10 \times 18 = 180$$

Since one numerical value 12 was misread as 22, therefore after subtracting the incorrect value and then adding the correct value in the total $n \bar{x}$, we have $180 - 22 + 12 = 170$. Hence, correct mean is $\bar{x} = \sum x/n = 170/10 = 17$ per cent.

(b) Given that $n = 200$, $\bar{x} = 50$. We know that

$$\bar{x} = \frac{\sum x}{n} \quad \text{or} \quad \sum x = n \bar{x} = 200 \times 50 = 10,000$$

Since two observations were misread, therefore the correct total $\sum x = n \bar{x}$ can be obtained as:

$$\sum x = 10,000 - (92 + 8) + (192 + 88) = 10,180$$

$$\text{Hence, correct mean is : } \bar{x} = \frac{\sum x}{n} = \frac{10,180}{200} = 50.9$$

Case 5: Frequency Distributions have Open-Ended Class Intervals

Example 3.17: The annual salaries (in rupees thousands) of employees in an organization are given below: The total salary of 10 employees in the class over Rs 40,000 is Rs 9,00,000. Compute the mean salary. Every employee belonging to the top 25 per cent of earners has to pay 5 per cent of his salary to the workers' relief fund. Estimate the contribution to this fund.

Salary (Rs '000)	Number of Employees
below 10	4
10–20	6
20–30	10
30–40	20
40 and above	10

Solution: Since class intervals are uniform, therefore we can take some width for open-end class intervals also. Calculations of mean are shown in Table 3.16

Table 3.16

Salary (Rs '000)	Mid-value (m_i)	Number of Employees (f_i)	$d_i = \frac{m_i - 25}{10}$	$f_i d_i$
0–10	5	4	-2	-8
10–20	15	6	-1	-6
20–30	25 ← A	10	0	0
30–40	35	20	1	20
40 and above	45 (given)	10	2	20
		50		26

where mid-value 25 is considered as the assumed mean. Applying the formula, we get

$$\bar{x} = A + \frac{\sum f_i d_i}{n} \times h = 25 + \frac{26}{50} \times 10 = \text{Rs } 30.2$$

The number of employees belonging to the top 25 per cent of the earners are $0.25 \times 50 = 13$ employees and the distribution of these top earners would be as follows:

Salary (Rs '000)	Number of Employees
40 and above	10
30–40	3

This calculation implies that 3 employees have been selected from the salary range 30–40. Under the assumption that frequencies are equally distributed between lower and upper limits of a class interval, the calculations would be as follows:

Since 20 employees have salary in the range 30–40 = 10 or Rs 10,000, therefore 3 employees will have income in the range $(10/20) \times 3 = 1.5$ or Rs 1,500. But we are interested in the top 3 earners in the range 30–40, their salaries will range from $(40 - 1.5)$ to 40, i.e., 38.5 to 40. Thus, the distribution of salaries of the top 25 persons is as follows:

Salary (Rs '000)	Mid-value (m_i)	Number of Employees (f_i)	Total Salary ($m_i f_i$)
40 and above	—	10	9,00,000 (given)
30–40	35	3	1,05,000
		13	10,05,000

This shows that the total income of the top 25 per cent of earners is Rs 10,05,000. Hence 5 per cent contribution to the fund is $0.05 \times 10,05,000 = \text{Rs } 50,250$.

Remark: If the width of class intervals is not same, then in accordance with the magnitude of change in the width, fix the width of last class interval.

3.5.4 Advantages and Disadvantages of Arithmetic Mean

Advantages

- The calculation of arithmetic mean is simple and it is unique, that is, every data set has one and only one mean.

- (ii) The calculation of arithmetic mean is based on all values given in the data set.
- (iii) The arithmetic mean is reliable single value that reflects all values in the data set.
- (iv) The arithmetic mean is least affected by fluctuations in the sample size. In other words, its value, determined from various samples drawn from a population, varies by the least possible amount.
- (v) It can be readily put to algebraic treatment. Some of the algebraic properties of arithmetic mean are as follows:
 - (a) *The algebraic sum of deviations of all the observations x_i ($i = 1, 2 \dots, n$) from the A.M. is always zero, that is,*

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = \sum_{i=1}^n x_i - n\left(\frac{1}{n}\right) \sum_{i=1}^n x_i = 0$$

Here the difference $x_i - \bar{x}$ ($i = 1, 2, \dots, n$) is usually referred to as *deviation from the arithmetic mean*. This result is also true for grouped data.

Due to this property, the mean is characterized as a *point of balance*, i.e. sum of the positive deviations from mean is equal to the sum of the negative deviations from mean.

- (b) *The sum of the squares of the deviations of all the observations from the A.M. is less than the sum of the squares of all the observations from any other quantity.*

Let x_i ($i = 1, 2, \dots, n$) be the given observations and \bar{x} be their arithmetic mean. Then this property implies that

$$\sum_{i=1}^n (x_i - \bar{x})^2 \leq \sum_{i=1}^n (x_i - a)^2$$

where 'a' is any constant quantity.

This property of A.M. is also known as the *least square property* and shall be quite helpful in defining the concept of standard deviation.

- (c) *It is possible to calculate the combined (or pooled) arithmetic mean of two or more than two sets of data of the same nature.*

Let \bar{x}_1 and \bar{x}_2 be arithmetic means of two sets of data of the same nature, of size n_1 and n_2 respectively. Then their *combined A.M.* can be calculated as:

$$\bar{x}_{12} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} \quad (3-7)$$

The result (3-7) can also be generalized in the same way for more than two sets of data of different sizes having different arithmetic means.

- (d) While compiling the data for calculating arithmetic mean, it is possible that we may wrongly read and/or write certain number of observations. In such a case, the correct value of A.M. can be calculated first by subtracting the sum of observations wrongly recorded from Σx_i (total of all observations) and then adding the sum of the correct observations to it. The result is then divided by the total number of observations.

Disadvantages

- (i) The value of A.M. cannot be calculated accurately for unequal and open-ended class intervals either at the beginning or end of the given frequency distribution.
- (ii) The A.M. is reliable and reflects all the values in the data set. However, it is very much affected by the extreme observations (or outliers) which are not representative of the rest of the data set. Outliers at the high end will increase the mean, while outliers at the lower end will decrease it. For example, if monthly income of four persons is 50, 70, 80, and 1000, then their A.M. will be 300, which does not represent the data.
- (iii) The calculation of A.M. sometime becomes difficult because every data element is used in the calculation (unless the short cut method for grouped data is used to calculate the mean). Moreover, the value so obtained may not be among the observations included in the data.

- (iv) The mean cannot be calculated for qualitative characteristics such as intelligence, honesty, beauty, or loyalty.
- (v) The mean cannot be calculated for a data set that has open-ended classes at either the high or low end of the scale.

Example 3.18: The mean salary paid to 1500 employees of an organization was found to be Rs 12,500. Later on, after disbursement of salary, it was discovered that the salary of two employees was wrongly entered as Rs 15,760 and 9590. Their correct salaries were Rs 17,760 and 8590. Calculate correct mean.

Solution: Let x_i ($i = 1, 2, \dots, 1500$) be the salary of i th employee. Then we are given that

$$\bar{x} = \frac{1}{1500} \sum_{i=1}^{1500} x_i = 12,500$$

$$\text{or } \sum_{i=1}^{1500} x_i = 12,500 \times 1500 = \text{Rs } 1,87,50,000$$

This gives the total salary disbursed to all 1500 employees. Now after adding the correct salary figures of two employees and subtracting the wrong salary figures posted against two employees, we have

$$\begin{aligned} \sum x_i &= 1,87,50,000 + (\text{Sum of correct salaries figures}) \\ &\quad - (\text{Sum of wrong salaries figures}) \\ &= 18,75,000 + (17,760 + 8590) - (15,760 + 9590) \\ &= 1,87,50,000 + 26,350 - 25,350 = 1,88,01,700 \end{aligned}$$

Thus the correct mean salary is given by

$$\bar{x} = 1,88,01,700 \div 1500 = \text{Rs } 12,534.46$$

Example 3.19: There are two units of an automobile company in two different cities employing 760 and 800 persons, respectively. The arithmetic means of monthly salaries paid to persons in these two units are Rs 18,750 and Rs 16,950 respectively. Find the combined arithmetic mean of salaries of the employees in both the units.

Solution: Let n_1 and n_2 be the number of persons working in unit 1 and 2 respectively, and \bar{x}_1 and \bar{x}_2 be the arithmetic mean of salaries paid to these persons respectively. We are given that:

$$\text{Unit 1: } n_1 = 760 ; \bar{x}_1 = \text{Rs } 18,750$$

$$\text{Unit 2: } n_2 = 800 ; \bar{x}_2 = \text{Rs } 16,950$$

Thus the combined mean of salaries paid by the company is:

$$\begin{aligned} \bar{x}_{12} &= \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{760 \times 18750 + 800 \times 16950}{760 + 800} \\ &= \text{Rs } 17,826.92 \text{ per month} \end{aligned}$$

Example 3.20: The mean yearly salary paid to all employees in a company was Rs 24,00,000. The mean yearly salaries paid to male and female employees were Rs 25,00,000 and Rs 19,00,000, respectively. Find out the percentage of male to female employees in the company.

Solution: Let n_1 and n_2 be the number of employees as male and female, respectively. We are given that

Characteristics	Groups		Combined Group (Total Employees)
	Male	Female	
Number of employees	$n_1 = ?$	$n_2 = ?$	$n = n_1 + n_2$
Mean salary (Rs)	$\bar{x}_1 = 25,00,000$	$\bar{x}_2 = 19,00,000$	$\bar{x}_{12} = 24,00,000$

Applying the formula for mean of combined group:

$$\bar{x}_{12} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

or

$$(n_1 + n_2) \bar{x}_{12} = n_1 \bar{x}_1 + n_2 \bar{x}_2$$

$$(n_1 + n_2) 24,00,000 = 25,00,000 n_1 + 19,00,000 n_2$$

$$1000 n_1 = 5000 n_2$$

$$\frac{n_1}{n_2} = \frac{5000}{1000} = \frac{5}{1}$$

$$n_1 : n_2 = 5 : 1$$

Hence, male employees in the company are $\{5 \div (5 + 1)\} \times 100 = 83.33$ per cent and female employees are $\{1 \div (5 + 1)\} \times 100 = 16.67$ per cent.

3.5.5 Weighted Arithmetic Mean

The arithmetic mean, as discussed earlier, gives equal importance (or weight) to each observation in the data set. However, there are situations in which values of individual observations in the data set are not of equal importance. If values occur with different frequencies, then computing A.M. of values (as opposed to the A.M. of observations) may not be a true representative of the data set characteristic and thus may be misleading. Under these circumstances, we may attach to each observation value a ‘weight’ w_1, w_2, \dots, w_N as an indicator of their importance perhaps because of size or importance and compute a weighted mean or average denoted by \bar{x}_w as:

$$\mu_w \text{ or } \bar{x}_w = \frac{\sum x_i w_i}{\sum w_i}$$

This is similar to the method for dealing with frequency data when the value is multiplied by the frequency, within each class, totalled and divided by the total number of values.

Remark: The **weighted arithmetic mean** should be used

- (i) when the importance of all the numerical values in the given data set is not equal.
- (ii) when the frequencies of various classes are widely varying
- (iii) where there is a change either in the proportion of numerical values or in the proportion of their frequencies.
- (iv) when ratios, percentages, or rates are being averaged.

Weighted arithmetic mean

mean: The mean for a data set obtained by assigning each observation a weight that reflects its importance within the data set.

Example 3.21: An examination was held to decide the awarding of a scholarship. The weights of various subjects were different. The marks obtained by 3 candidates (out of 100 in each subject) are given below:

Subject	Weight	Students		
		A	B	C
Mathematics	4	60	57	62
Physics	3	62	61	67
Chemistry	2	55	53	60
English	1	67	77	49

Calculate the weighted A.M. to award the scholarship.

Solution: The calculations of the weighted arithmetic mean is shown in Table 3.17

Table 3.17 Calculations of Weighted Arithmetic Mean

Subject	Weight (w_i)	Students					
		Student A		Student B		Student C	
		Marks (x_i)	$x_i w_i$	Marks (x_i)	$x_i w_i$	Marks (x_i)	$x_i w_i$
Mathematics	4	60	240	57	228	62	248
Physics	3	62	186	61	183	67	201
Chemistry	2	55	110	53	106	60	120
English	1	67	67	77	77	49	49
	10	244	603	248	594	238	618

Applying the formula for weighted mean, we get

$$\bar{x}_{wA} = \frac{603}{10} = 60.3 ; \quad \bar{x}_A = \frac{244}{4} = 61$$

$$\bar{x}_{wB} = \frac{594}{10} = 59.4 ; \quad \bar{x}_B = \frac{248}{4} = 62$$

$$\bar{x}_{wC} = \frac{618}{10} = 61.8 ; \quad \bar{x}_c = \bar{x}_e = 59.5$$

From the above calculations, it may be noted that student B should get the scholarship as per simple A.M. values, but according to weighted A.M., student C should get the scholarship because all the subjects of examination are not of equal importance.

Example 3.22: The owner of a general store was interested in knowing the mean contribution (sales price minus variable cost) of his stock of 5 items. The data is given below:

Product	Contribution per Unit	Quantity Sold
1	6	160
2	11	60
3	8	260
4	4	460
5	14	110

Solution: If the owner ignores the values of the individual products and gives equal importance to each product, then the mean contribution per unit sold will be

$$\bar{x} = (1 \div 5) \{6 + 11 + 8 + 4 + 14\} = \text{Rs } 8.6$$

This value, Rs 8.60 may not necessarily be the mean contribution per unit of different quantities of the products sold. In this case the owner has to take into consideration the number of units of each product sold as different weights. Computing weighted A.M. by multiplying units sold (w) of a product by its contribution (x). That is,

$$\bar{x}_w = \frac{6(160) + 11(60) + 8(260) + 4(460) + 14(110)}{160 + 60 + 260 + 460 + 110} = \frac{7,080}{1,050} = \text{Rs } 6.74$$

This value, Rs 6.74, is different from the earlier value, Rs 8.60. The owner must use the value Rs 6.74 for decision-making purpose.

Example 3.23: A management consulting firm, has four types of professionals on its staff: managing consultants, senior associates, field staff, and office staff. Average rates charged to consulting clients for the work of each of these professional categories are Rs 3150/hour, Rs 1680/hour, Rs 1260/hour, and 630/hour respectively. Office records indicate the following number of hours billed last year in each category: 8000, 14,000, 24,000, and 35,000 respectively. If the firm is trying to come up with an average billing rate for estimating client charges for next year, what would you suggest they do and what do you think is an appropriate rate?

Solution: The data given in the problem are as follows:

Staff	Consulting Charges (Rs per hour)	Hours Billed
	x_i	w_i
Managing consultants	3150	8000
Senior associates	1680	14,000
Field staff	1260	24,000
Office staff	630	35,000

Applying the formula for weighted mean, we get,

$$\begin{aligned}\bar{x}_w &= \frac{\sum x_i w_i}{\sum w_i} = \frac{3150(8000) + 1680(14,000) + 1260(24,000) + 630(35,000)}{8000 + 14,000 + 24,000 + 35,000} \\ &= \frac{2,52,00,000 + 2,35,20,000 + 3,02,40,000 + 2,20,50,000}{81,000} \\ &= \text{Rs } 1247.037 \text{ per hour}\end{aligned}$$

However, the firm should cite this as an average rate for clients who use the four professional categories for approximately 10 per cent, 17 per cent, 30 per cent and 43 per cent of the total hours billed.

Conceptual Questions 3A

- Explain the term *average*. What are the merits of a good average? Explain with examples.
- What are the measures of central tendency? Why are they called measures of central tendency?
- What are the different measures of central tendency? Mention the advantages and disadvantages of arithmetic mean.
- What are the different measures of central tendency? Discuss the essentials of an ideal average.
- Give a brief description of the various measures of central tendency. Why is arithmetic mean so popular?
- What information about a body of data is provided by an average? How are averages useful as a descriptive measure?
- It is said that the weighted mean is commonly referred to as a ‘weighted average’. How is the use of this phrase inconsistent with the definition of an average?
- How is an average considered as a representative measure or a measure of central tendency? How is the ability of an average to measure central tendency related to other characteristics of data?
- Prove that the algebraic sum of the deviations of a given set of observations from their arithmetic mean is zero.
[MBA, UP Tech. Univ., 2000]
- Is it necessarily true that being above average indicates that someone is superior? Explain.
[Delhi Univ., MBA, 2000]
- What is statistical average? What are the desirable properties for an average to possess? Mention the different types of averages and state why arithmetic mean is the most commonly used amongst them.
- Distinguish between simple and weighted average and state the circumstances under which the latter should be employed.

Self-Practice Problems 3A

- An investor buys Rs 12,000 worth of shares of a company each month. During the first 5 months he bought the shares at a price of Rs 100, Rs 120, Rs 150, Rs 200, and Rs 240 per share respectively. After 5 months what is the average price paid for the shares by him?
- A company wants to pay bonus to members of the staff. The bonus is to be paid as under:

Monthly Salary (in Rs)	Bonus
3000–4000	1000
4000–5000	1200
5000–6000	1400
6000–7000	1600
7000–8000	1800
8000–9000	2200
9000–10,000	2200
10,000–11,000	2400

Actual amount of salary drawn by the employees is given below:

3250	3780	4200	4550	6200	6600
6800	7250	3630	8320	9420	9520
8000	10,020	10,280	11,000	6100	6250
7630	3820	5400	4630	5780	7230
	6900				

How much would the company need to pay by way of bonus? What shall be the average bonus paid per member of the staff?

- 3.3** Calculate the simple and weighted arithmetic mean price per tonne of coal purchased by a company for the half year. Account for difference between the two:

Month	Price/tonne	Tonnes Purchased	Month	Price/tonne	Tonnes Purchased
January	4205	25	April	5200	52
February	5125	30	May	4425	10
March	5000	40	June	5400	45

- 3.4** Salary paid by a company to its employees is as follows:

Designation	Monthly Salary (in Rs)	Number of Persons
Senior Manager	35,000	1
Manager	30,000	20
Executives	25,000	70
Jr Executives	20,000	10
Supervisors	15,000	150

Calculate the simple and weighted arithmetic mean of salary paid.

- 3.5** The capital structure of a company is as follows:

	Book Value (Rs)	After Tax Cost (%)
Equity	2,15,00,000	19
Preference share	2,07,00,000	11
Debt	2,11,00,000	9

Calculate the weighted average cost of capital.

- 3.6** The mean monthly salary paid to all employees in a company is Rs 16,000. The mean monthly salaries paid to technical and non-technical employees are Rs 18,000 and Rs 12,000 respectively. Determine the percentage of technical and non-technical employees in the company.
- 3.7** The mean marks in statistics of 100 students in a class was 72 per cent. The mean marks of boys was 75 per cent, while their number was 70 per cent. Find out the mean marks of girls in the class.

- 3.8** Mr. Gupta, a readymade garments store owner, advertises: 'If our average prices are not equal or lower than everyone else's, you get it free.' One customer came into the store one day and threw on the counter bills of six items he had bought from a competitor for an average price less than Gupta's. The items cost (in Rs): 201.29 202.97 203.49 205.00 207.50 210.95

The prices for the same six items at Mr. Gupta's stores

are (in Rs): 201.35, 202.89, 203.19, 204.98, 207.59 and 211.50. Mr. Gupta told the customer, My advertisement refers to a weighted average price of these items. Our average is lower because our sales of these items have been: 207, 209, 212, 208, 206, and 203 (in units).

Is Mr. Gupta getting himself into or out of trouble with his contention about weighted average?

- 3.9** The arithmetic mean height of 50 students of a college is 5'8". The height of 30 of these is given in the frequency distribution below. Find the arithmetic mean height of the remaining 20 students.

Height in inches : 5'4" 5'6" 5'8" 5'10" 6'0"

Frequency : 4 12 4 8 2

- 3.10** The following table gives salary per month of 450 employees in a factory:

Salary	No. of Employees
Less than 5000	80
5000–10,000	120
10,000–15,000	100
15,000–20,000	60
20,000–25,000	50
25,000–30,000	40

The total income of 6 persons in the group 25,000–30,000 is Rs 1,65,000. Due to a rise in prices, the factory owner decided to give adhoc increase of 25 per cent of the average pay to the 25 per cent of the lowest paid employees, 10 per cent of the average pay to the 10 per cent highest paid employees and 15 per cent to the remaining employees.

Find out the additional amount required for the adhoc increase and after the increase, find out the average pay of an employee in the factory.

- 3.11** A professor of management has decided to use weighted average to find the internal assessment grades of his students on the basis of following parameters: Quizzes—30 per cent, Term Paper—25 per cent, Mid-term test—30 per cent and Class attendance—15 per cent. From the data below, compute the final average in the internal assessment

Student	Quizzes	Term Paper	Mid-Term	Attendance
1	55	59	64	20
2	48	54	58	22
3	64	58	63	19
4	52	49	58	23
5	65	60	62	18

- 3.12** An appliances manufacturing company is forecasting regional sales for next year. The Delhi branch, with current yearly sales of Rs 387.6 million, is expected to achieve a sales growth of 7.25 percent; the Kolkata branch, with current sales of Rs 158.6 million, is expected to grow by 8.20 per cent; and the Mumbai branch, with sales of Rs 115 million, is expected to increase sales by 7.15 per cent. What is the average rate of growth forecasted for next year?

Hints and Answers

3.1 $x = \text{Rs } 146.30$

3.2 Rs 42,000 to pay bonus; Average bonus paid per member $= 42,000/25 = \text{Rs } 1,680$

3.3 $\bar{x} = \text{Rs } 4892.5 \text{ tonne}$; $\bar{x}_w = 5032.30$

3.4 $\bar{x} = \text{Rs } 25,000$; $\bar{x}_w = \text{Rs } 19,262.94$

3.6 Percentage of technical personnel $= 66.67$ per cent ; Non-technical $= 33.33$ per cent

3.7 Mean marks of girls, $\bar{x}_2 = 65$ per cent

3.9 \bar{x}_1 (mean height of $n_1 = 30$ students) $= 5'6''$.

Given $n_2 = 50 - 30 = 20$, \bar{x} (mean height of 50 students) $= 68''$. Thus

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = 5'9''$$

3.10 \bar{x} (average wage) $= \text{Rs } 17,870$;

Given $n_1 = 25\%$ lowest paid employee $500/4 = 125$

\bar{x}_1 = Average salary to these 25 per cent employees $= \text{Rs } 17,870 + 17,870/4$

$$= \text{Rs } 22,337.50$$

$n_2 = 20\%$ highest paid employees $= 500/10 = 50$

\bar{x}_2 = Average salary of these 20% employees $= 17,870 + 17,870/10 = \text{Rs } 19,657.00$

$$n_3 = 500 - (125 + 50) = 325$$

$$\begin{aligned}\bar{x}_3 &= \text{Average salary to these employees} \\ &= 17,870 + (15 \times 17,870)/100 \\ &= \text{Rs } 20,550.50\end{aligned}$$

$$\therefore \bar{x}_{123} = \text{Rs } 2,090.90.$$

- 3.11** Student 1 : $0.30 \times 55 + 0.15 \times 59 + 0.30 \times 64 + 0.15 \times 20 = 16.5 + 8.85 + 19.2 + 3.0 = 47.55$
 2 : $0.20 \times 48 + 0.15 \times 54 + 0.30 \times 58 + 0.15 \times 22 = 14.4 + 8.10 + 17.4 + 3.3 = 43.20$
 3: $0.30 \times 64 + 0.15 \times 58 + 0.30 \times 63 + 0.15 \times 19 = 19.2 + 8.7 + 18.9 + 2.85 = 49.65$
 4: $0.30 \times 52 + 0.15 \times 49 + 0.30 \times 58 + 0.15 \times 23 = 15.6 + 7.35 + 17.4 + 3.45 = 43.80$
 5: $0.30 \times 65 + 0.15 \times 60 + 0.30 \times 62 + 0.15 \times 18 = 19.5 + 9.0 + 18.6 + 2.7 = 49.8$

$$\begin{aligned}\text{3.12 } \bar{x}_w &= \frac{\sum x_i w_i}{\sum w_i} \\ &= \frac{387.6 \times 7.25 + 158.6 \times 8.20 + 115 \times 7.15}{387.6 + 158.6 + 115} \\ &= \frac{2810.10 + 1300.52 + 822.25}{661.20} \\ &= \frac{4932.87}{661.20} = 7.46 \text{ per cent}\end{aligned}$$

3.6 GEOMETRIC MEAN

In many business and economics problems, we deal with quantities (variables) that change over a period of time. In such cases the aim is to know an average percentage change rather than simple average value to represent the average growth or declining rate in the variable value over a period of time. Thus we need to calculate another measure of central tendency called **geometric mean** (G.M.). The specific application of G.M. is to show multiplicative effects over time in compound interest and inflation calculations.

Consider, for example, the annual rate of growth of output of a company in the last five years.

Geometric mean: A value that represents n th root of the product of a set of n numbers.

Year	Growth Rate (Per cent)	Output at the End of the Year
1998	5.0	105
1999	7.5	112.87
2000	2.5	115.69
2001	5.0	121.47
2002	10.0	133.61

The simple arithmetic mean of the growth rate is:

$$\bar{x} = \frac{1}{5}(5 + 7.5 + 2.5 + 5 + 10) = 6$$

This value of 'mean' implies that if 6 per cent is the growth rate, then output at the end of 2002 should be 133.81, which is slightly more than the actual value, 133.61. Thus the correct growth rate should be slightly less than 6.

To find the correct growth rate, we apply the formula of geometric mean:

$$\begin{aligned} \text{G.M.} &= \sqrt[n]{\text{Product of all the } n \text{ values}} \\ &= \sqrt[n]{x_1 \cdot x_2 \cdots x_n} = (x_1 \cdot x_2 \cdots x_n)^{\frac{1}{n}} \end{aligned} \quad (3-9)$$

In other words, *G.M. of a set of n observations is the nth root of their product.*

For the above example, substituting the values of growth rate in the given formula, we have

$$\begin{aligned} \text{G.M.} &= \sqrt[5]{5 \times 7.5 \times 2.5 \times 5 \times 10} = \sqrt[5]{4687.5} \\ &= 5.9 \text{ per cent average growth.} \end{aligned}$$

Calculation of G.M.

When the number of observations are more than three, the G.M. can be calculated by taking logarithm on both sides of the equation. The formula (3-9) for G.M. for ungrouped data can be expressed in terms of logarithm as shown below:

$$\begin{aligned} \text{Log (G.M.)} &= \frac{1}{n} \log (x_1 \cdot x_2 \cdots x_n) \\ &= \frac{1}{n} \{ \log x_1 + \log x_2 + \dots + \log x_n \} = \frac{1}{n} \sum_{i=1}^n \log x_i \end{aligned}$$

and therefore

$$\text{G.M.} = \text{antilog} \left\{ \frac{1}{n} \sum \log x_i \right\} \quad (3-10)$$

If the observations x_1, x_2, \dots, x_n occur with frequencies f_1, f_2, \dots, f_n , respectively, and the total of frequencies is $n = \sum f_i$, then the G.M. for such data is given by

$$\begin{aligned} \text{G.M.} &= \left(x_1^{f_1} \cdot x_2^{f_2} \cdots x_n^{f_n} \right)^{1/n} \\ \text{or} \quad \log (\text{G.M.}) &= \frac{1}{n} \{ f_1 \log x_1 + f_2 \log x_2 + \dots + f_n \log x_n \} \\ &= \frac{1}{n} \sum_{i=1}^n f_i \log x_i \\ \text{or} \quad \text{G.M.} &= \text{Antilog} \left\{ \frac{1}{n} \sum f_i \log x_i \right\} \end{aligned} \quad (3-11)$$

Example 3.24: The rate of increase in population of a country during the last three decades is 5 per cent, 8 per cent, and 12 per cent. Find the average rate of growth during the last three decades.

Solution: Since the data is given in terms of percentage, therefore geometric mean is a more appropriate measure. The calculations of geometric mean are shown in Table 3.18:

Table 3.18 Calculations of G.M.

Decade	Rate of Increase in Population (%)	Population at the End of Decade (x) Taking Preceding Decade as 100	$\log_{10} x$
1	5	105	2.0212
2	8	108	2.0334
3	12	112	2.0492

Using the formula (3-10), we have

$$\begin{aligned} \text{G.M.} &= \text{Antilog} \left\{ \frac{1}{n} \sum \log x \right\} = \text{Antilog} \left\{ \frac{1}{3} (6.1038) \right\} \\ &= \text{Antilog} (2.0346) = 108.2 \end{aligned}$$

Hence the average rate of increase in population over the last three decades is $108.2 - 100 = 8.2$ per cent.

Example 3.25: A given machine is assumed to depreciate 40 per cent in value in the first year, 25 per cent in the second year, and 10 per cent per year for the next three years, each percentage being calculated on the diminishing value. What is the average depreciation recorded on the diminishing value for the period of five years?

Solution: The calculation of geometric mean is shown in Table 3.19.

Table 3.19 Calculation of G.M.

Rate of Depreciation (x_i) (in percentage)	Number of Years (f_i)	$\log_{10} x_i$	$f_i \log_{10} x_i$
40	1	1.6021	1.6021
25	1	1.3979	1.3979
10	3	1.0000	3.0000
			6.0000

Using formula (3-11), we have

$$\begin{aligned} \text{G.M.} &= \text{Antilog} \left\{ \frac{1}{n} \sum f \log x \right\} = \text{Antilog} \left\{ \frac{1}{5} (6.0000) \right\} \\ &= \text{Antilog} (1.2) = 15.85 \end{aligned}$$

Hence, the average rate of depreciation for first five years is 15.85 per cent.

3.6.1 Combined Geometric Mean

The combined geometric mean of observations formed by pooling the geometric means of different sets of data is defined as:

$$\log \text{G.M.} = \frac{\sum_{i=1}^n n_i \log G_i}{\sum_{i=1}^n n_i} \quad (3-12)$$

where G_i is the geometric mean of the i th data set having n_i number of observations.

3.6.2 Weighted Geometric Mean

If different observations x_i ($i = 1, 2, \dots, n$) are given different weights (importance), say w_i ($i = 1, 2, \dots, n$) respectively, then their weighted geometric mean is defined as:

$$\begin{aligned} \text{G.M.}(w) &= \text{Antilog} \left[\left(\frac{1}{n} \right) \sum w \log x \right] \\ &= \text{Antilog} \left[\left(\frac{1}{\sum w} \right) \sum w \log x \right] \quad (3-13) \end{aligned}$$

Example 3.26: Three sets of data contain 8, 7, and 5 observations and their geometric means are 8.52, 10.12, and 7.75, respectively. Find the combined geometric mean of 20 observations.

Solution: Applying the formula (3-12), the combined geometric mean can be obtained as follows:

$$\begin{aligned}
 \text{G.M.} &= \text{Antilog} \left[\frac{n_1 \log G_1 + n_2 \log G_2 + n_3 \log G_3}{n_1 + n_2 + n_3} \right] \\
 &= \text{Antilog} \left[\frac{8 \log (8.52) + 7 \log (10.12) + 5 \log (7.75)}{8 + 7 + 5} \right] \\
 &= \text{Antilog} \left[\frac{(8 \times 0.9304) + (7 \times 1.0051) + (5 \times 0.8893)}{20} \right] \\
 &= \text{Antilog} \left(\frac{18.9254}{20} \right) = \text{Antilog} (0.94627) = 8.835
 \end{aligned}$$

Hence, the combined G.M. of 20 observations is 8.835.

Example 3.27: The weighted geometric mean of four numbers 8, 25, 17, and 30 is 15.3. If the weights of the first three numbers are 5, 3, and 4 respectively, find the weight of fourth number.

Solution: Let weight of fourth number be w . Then the weighted geometric mean of four numbers can be calculated as shown in Table 3.20.

Table 3.20 Calculations of Weighted G.M.

Numbers (x)	Weight of Each Number (w)	$\log_{10} x$	$w \log_{10} x$
8	5	0.9031	4.5155
25	3	1.3979	4.1937
17	4	1.2304	4.9216
30	w	1.4771	1.4771 w
		12 + w	13.6308 + 1.4771 w

Thus the weighted G.M. is

$$\begin{aligned}
 \log \{\text{G.M.} (w)\} &= \left[\left(\frac{1}{\sum w} \right) \sum w \log x \right] \\
 \text{or} \quad \log (15.3) &= \left[\left(\frac{1}{12+w} \right) (13.6308 + 1.4771w) \right]
 \end{aligned}$$

$$(1.1847)(12 + w) = 13.6308 + 1.4771w$$

$$14.2164 + 1.1847w = 13.6308 + 1.4771w$$

$$0.5856 = 0.2924 w$$

$$\text{or} \quad w = \frac{0.5856}{0.2924} = 2 \text{ (approx.)}$$

Thus the weight of fourth number is 2.

3.6.3 Advantages, Disadvantages, and Applications of G.M.

Advantages

- (i) The value of G.M. is not much affected by extreme observations and is computed by taking all the observations into account.
- (ii) It is useful for averaging ratio and percentage as well as in determining rate of increase and decrease.
- (iii) In the calculation of G.M. more weight is given to smaller values and less weight to higher values. For example, it is useful in the study of price fluctuations where the lower limit can touch zero whereas the upper limit may go upto any number.

- (iv) It is suitable for algebraic manipulations. The calculation of weighted G.M. and combined G.M. are two examples of algebraic manipulations of the original formula of geometric mean.

Disadvantages

- (i) The calculation of G.M. as compared to A.M., is more difficult and intricate.
- (ii) The value of G.M. cannot be calculated when any of the observations in the data set is either negative or zero.
- (iii) While calculating weighted geometric, mean equal importance (or weight) is not given to each observation in the data set.

Applications

- (i) The concept of G.M. is used in the construction of index numbers.
- (ii) Since $G.M. \leq A.M.$, therefore G.M. is useful in those cases where smaller observations are to be given importance. Such cases usually occur in social and economic areas of study.
- (iii) The G.M. of a data set is useful in estimating the average rate of growth in the initial value of an observation per unit per period. For example, it is useful in finding the percentage increase in sales, profit, production, population, and so on. It is also useful in calculating the amount of money accumulated at the end of n periods, with an original principal amount of P_0 . The formula is as follows:

$$P_n = P_0 (1 + r)^n$$

$$\text{or } r = \left(\frac{P_n}{P_0} \right)^{\frac{1}{n}} - 1$$

where r = interest rate (rate of growth) per unit period
 n = number of years or length of the period.

Conceptual Questions 3B

14. Define simple and weighted geometric mean of a given distribution. Under what circumstances would you recommend its use?
15. Discuss the advantages, disadvantages, and uses of geometric mean.

Self-Practice Problems 3B

- 3.13** Find the geometric mean of the following distribution of data:

Dividend declared (%) : 0–10 10–20 20–30 30–40 40–45

Number of companies : 5 7 15 25 8

- 3.14** The population of a country was 300 million in 1985. It became 520 million in 1995. Calculate the percentage compounded rate of growth per year.

- 3.15** Compared to the previous year, the overhead expenses went up by 32 per cent in 1994, increased by 40 per cent in the next year, and by 50 per cent in the following year. Calculate the average rate of increase in overhead expenses over the three years.

- 3.16** The rise in the price of a certain commodity was 5 per cent in 1995, 8 per cent in 1996, and 77 per cent in

1997. It is said that the average price rise between 1995 and 1997 was 26 per cent and not 30 per cent. Justify the statement and show how you would explain it before a layman.

- 3.17** The weighted geometric mean of the four numbers 20, 18, 12, and 4 is 11.75. If the weights of the first three numbers are 1, 3, and 4 respectively, find the weight of the fourth number.

- 3.18** A machinery is assumed to depreciate 44 per cent in value in the first year, 15 per cent in the second year, and 10 per cent per year for the next three years, each percentage being calculated on diminishing value. What is the average percentage of depreciation for the entire period?

- 3.19** The following data represent the percentage increase in the number of prisoners (a negative number indicates a percentage decrease) in a district jail:

Year	:	1995	1996	1997	1998	1999	2000
Per cent increase	:	-2	3	7	4	5	-3

Calculate the average percentage increase using data from 1996–1999 as well as using data for all 6 years.

- 3.20** A manufacturer of electrical circuit boards, has manufactured the following number of units over the past 5 years:

2000	2001	2002	2003	2004
14,300	15,150	16,110	17,540	19,430

Calculate the average percentage increase in units produced over this time period, and use this to estimate production for 2006.

- 3.21** The owner of a warehouse is calculating the average growth factor for his warehouse over the last 6 years. Using a geometric mean, he comes up with an answer of 1.42. Individual growth factors for the first 5 years were 1.91, 1.53, 1.32, 1.91, and 1.40, but he lost the records for the sixth year, after he calculated the mean. What was it?

Hints and Answers

3.13 G.M. = 25.64 per cent.

3.14 Apply the formula: $P_n = P_0 (1 + r)^n$; r is the rate of growth in population. Since $P_{1995} = 520$ and $P_{1985} = 300$ and $n = 10$, therefore $520 = 300 (1 + r)^{10}$ or $r = 3.2$ per cent.

3.15 Apply the formula $P_n = P_0 (1 + r)^3 = P_0 (1 + 0.32)(1 + 0.40)(1 + 0.50)$; P_0 = Overhead expenses in 1994: $(1 + r)^3 = 1.32 \times 1.40 \times 1.50$. Taking log and simplified, we get $r = 40.5$ per cent.

3.16 Average price rise = 26 per cent (G.M.) If we use A.M. and take the price in the base year as 100, then $(105 + 108 + 177)/3 = 130$ or 30 per cent is the average change per year. Then the price in 1995 would be 130, price in 1996 would be $130 + 30$ per cent increase on 130 = 169, and in the year 1997 it would be $169 + 30$ per cent increase on 130 = 219.7

$$\begin{aligned}\text{3.17} \quad \text{Apply } \log \text{G.M.} &= \frac{\sum w \log w}{\sum w} \text{ or } \log 11.75 \\ &= \frac{9.4974 + 0.6021w_4}{8 + w_4} \text{ or } w_4 = 0.850 \text{ (approx.).}\end{aligned}$$

3.18 Depreciation rate : 44 15 10 10 10
Diminishing value
taking 100 as base (x) : 56 85 90 90 90

3.22 Industrial Gas Supplier keeps records on the cost of processing a purchase order. Over the last 5 years, this cost has been Rs 355, 358, 361, 365 and 366. What has supplier's average percentage increase been over this period? If this average rate stays the same for 3 more years, what will cost supplier to process a purchase order at that time?

3.23 A sociologist has been studying the yearly changes in the number of convicts assigned to the largest correctional facility in the state. His data are expressed in terms of the percentage increase in the number of prisoners (a negative number indicates a percentage decrease). The sociologist's most recent data are as follows:

1999	2000	2001	2002	2003	2004
5%	6%	9%	4%	7%	6%

- (a) Calculate the average percentage increase using only the 1999–2002 data.
(b) A new penal code was passed in 1998. Previously, the prison population grew at a rate of about 2 percent per year. What seems to be the effect of the new penal code?

$$\begin{aligned}\text{Log } x: 1.7582 \ 1.9294 \ 1.9542 \ 1.9542 \ 1.9542 &= 9.5502 \\ \text{G.M.} &= \text{Antilog} (\Sigma \log x/N) = \text{Antilog} (9.5502/5) \\ &= \text{Antilog} (1.91004) \\ &= 81.28\end{aligned}$$

The diminishing value is Rs 81.28 and average depreciation is 18.72 per cent.

3.20 G.M.: $\sqrt[4]{19430/14300} = \sqrt[4]{1.3587} = 1.07964$. So the average increase is 7.96 per cent per year. In 2006, the estimated production will be $19430 (1.0796)^2 = 22,646$ units (approx.)

3.21 Since G.M.: $1.42 = x \times \sqrt[6]{1.91 \times 1.53 \times 1.32 \times 1.91 \times 1.40}$
 $x = (1.42)^6 / (1.91 \times 1.53 \times 1.32 \times 1.91 \times 1.40)$
 $= 8.195/10.988 = 0.7458$

3.22 G.M. = $\sqrt[4]{366/355} = \sqrt[4]{1.0309} = 1.00765$. So the average increase is 0.765 per cent per year. In three more years the estimated cost will be $366 (1.00765)^3 = \text{Rs } 757.600$

3.23 (a) G.M.: $\sqrt[4]{0.95 \times 1.06 \times 1.09 \times 1.04} = \sqrt[4]{2.5132} = 1.03364$. So the average rate of increase from 1999–2002 was 3.364 per cent per year.
(b) G.M.: $\sqrt[6]{0.95 \times 1.06 \times 1.09 \times 1.04 \times 1.07 \times 0.94} = \sqrt[6]{1.148156} = 1.01741$. So the new code appears to have slight effect on the rate of growth of convicts, which has decreased from 2 per cent to 1.741 per cent per year.

3.7 HARMONIC MEAN

The **harmonic mean** (H.M.) of a set of observations is defined as the reciprocal of the arithmetic mean of the reciprocal of the individual observations, that is,

$$\frac{1}{H.M.} = \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}$$

or

$$H.M. = \frac{n}{\sum_{i=1}^n \left(\frac{1}{x_i} \right)} \quad (\text{For ungrouped data}) \quad (3-14)$$

If f_1, f_2, \dots, f_n are the frequencies of observations x_1, x_2, \dots, x_n , then the harmonic mean is defined as:

$$H.M. = \frac{n}{\sum_{i=1}^n f_i \left(\frac{1}{x_i} \right)} \quad (\text{For grouped data}) \quad (3-15)$$

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

Harmonic mean: A value that is the reciprocal of the mean of the reciprocals of a set of numbers.

Example 3.28: An investor buys Rs 20,000 worth of shares of a company each month. During the first 3 months he bought the shares at a price of Rs 120, Rs 160, and Rs 210. After 3 months what is the average price paid by him for the shares?

Solution: Since the value of shares is changing after every one month, therefore the required average price per share is the harmonic mean of the prices paid in first three months.

$$\begin{aligned} H.M. &= \frac{3}{(1/120) + (1/160) + (1/210)} = \frac{3}{0.008 + 0.006 + 0.004} \\ &= 3/0.018 = \text{Rs } 166.66 \end{aligned}$$

Example 3.29: Find the harmonic mean of the following distribution of data

Dividend yield (per cent) :	2–6	6–10	10–14
Number of companies :	10	12	18

Solution: The calculation of harmonic mean is shown in Table 3.21.

Table 3.21 Calculation of H.M.

Class Intervals (Dividend yield)	Mid-value (m_i)	Number of Companies (frequency, f_i)	Reciprocal $\left(\frac{1}{m_i} \right)$	$f_i \left(\frac{1}{m_i} \right)$
2–6	4	10	1/4	2.5
6–10	8	12	1/8	1.5
10–14	12	18	1/12	1.5
$N = 40$				5.5

$$\text{The harmonic mean is: } H.M. = \frac{n}{\sum_{i=1}^3 f_i \left(\frac{1}{m_i} \right)} = \frac{40}{5.5} = 7.27$$

Hence the average dividend yield of 40 companies is 7.27 per cent.

3.7.1 Advantages, Disadvantages, and Applications of H.M.

Advantages

- (i) The H.M. of the given data set is also computed based on its every element.
- (ii) While calculating H.M., more weightage is given to smaller values in a data set

because in this case, the reciprocal of given values is taken for the calculation of H.M.

- (iii) The original formula of H.M. can be extended to accommodate further analysis of data by certain algebraic manipulations.

Disadvantages

- (i) The H.M. is not often used for analysing business problems.
- (ii) The H.M. of any data set cannot be calculated if it has negative and/or zero elements.
- (iii) The calculation of H.M. involves complicated calculations. For calculating the H.M. of a data set, the largest weight is given to smaller values of elements, therefore it does not represent the true characteristic of the data set.

Applications

The harmonic mean is particularly useful for computation of average rates and ratios. Such rates and ratios are generally used to express relations between two different types of measuring units that can be expressed reciprocally. For example, distance (in km), and time (in hours).

3.8 RELATIONSHIP AMONG A.M., G.M., AND H.M.

For any set of observations, its A.M., G.M., and H.M. are related to each other in the relationship

$$\text{A.M.} \geq \text{G.M.} \geq \text{H.M.}$$

The sign of '=' holds if and only if all the observations are identical.

If observations in a data set take the values $a, ar, ar^2, \dots, ar^{n-1}$, each with single frequency, then

$$(\text{G.M.})^2 = \text{A.M.} \times \text{H.M.}$$

Self-Practice Problems 3C

- 3.24** In a certain factory, a unit of work is completed by A in 4 minutes, by B in 5 minutes, by C in 6 minutes, by D in 10 minutes, and by E in 12 minutes (a) What is the average rate of completing the work? (b) What is the average number of units of work completed per minute? (c) At this rate how many units will they complete in a six-hour day?
- 3.25** An investor buys Rs 12,000 worth of shares of a company each month. During the first 5 months he bought the shares at a price of Rs 100, Rs 120, Rs 150, Rs 200, and Rs 240 per share respectively. After 5 months what is the average price paid by him for the shares?

- 3.26** Calculate the A.M., G.M., and H.M. of the following observations and show that $\text{A.M.} > \text{G.M.} > \text{H.M.}$

32 35 36 37 39 41 43

- 3.27** The profit earned by 18 companies is given below:

Profit (in Rs lakh) : 20 21 22 23 24 25

No. of companies : 4 2 7 1 3 1

Calculate the harmonic mean of profit earned.

- 3.28** Find the harmonic mean for the following distribution of data:

Class interval : 0–10 10–20 20–30 30–40

Frequency : 5 8 3 4

Hints and Answers

- 3.24** (a) Average rate of completing the work per minute = 6.25 (b) Average units/minute = $1 \div 6.25 = 0.16$; (c) Units completed in six-hours (360 minutes) day by all 5 workers = $360 \times 0.16 = 288$ units

- 3.25** Average price paid for shares = Rs 146.30

- 3.26** A.M. = 37.56; G.M. = 37.52; H.M. = 37.25

- 3.27** H.M. = Rs 21.9 lakh

- 3.28** H.M. = 9.09

3.9 AVERAGES OF POSITION

Different from mathematical averages—arithmetic mean, geometric mean, and harmonic mean, which are mathematical in nature and deal with those characteristics of a data set which can be directly measured quantitatively, such as: income, profit, level of production, rate of growth, etc. However, in cases where we want to guard against the influence of a few outlying observations (called outliers), and/or we need to measure qualitative characteristics of a data set, such as: honesty, intelligence, beauty, consumer acceptance, and so on, other measures of central tendency namely *median*, *quartiles*, *deciles*, *percentiles*, and *mode* are used. These measures are also called *positional averages*. The term ‘position’ refers to the place of the value of an observation in the data set. These measures help in identifying the value of an observation of interest rather than computing it.

3.9.1 Median

Median may be defined as the *middle value* in the data set when its elements are arranged in a sequential order, that is, in either ascending or descending order of magnitude. It is called a middle value in an ordered sequence of data in the sense that half of the observations are smaller and half are larger than this value. The **median** is thus a measure of the *location* or *centrality* of the observations.

The median can be calculated for both ungrouped and grouped data sets.

Median: A measure of central location such that one half of the observations in the data set is less than or equal to the given value.

Ungrouped Data

In this case, the data is arranged in either ascending or descending order of magnitude.

- If the number of observations (n) is an *odd number*, then the median (Med) is represented by the numerical value corresponding to the positioning point of $(n + 1)/2$ ordered observation. That is,

$$\text{Med} = \text{Size or value of } \left(\frac{n+1}{2} \right) \text{th observation in the data array}$$

- If the number of observations (n) is an *even number*, then the median is defined as the arithmetic mean of the numerical values of $n/2$ th and $(n/2 + 1)$ th observations in the data array. That is,

$$\text{Med} = \frac{\frac{n}{2} \text{th observation} + \left(\frac{n}{2} + 1 \right) \text{th observation}}{2}$$

Example 3.30: Calculate the median of the following data that relates to the service time (in minutes) per customer for 7 customers at a railway reservation counter: 3.5, 4.5, 3, 3.8, 5.0, 5.5, 4

Solution: The data are arranged in ascending order as follows:

Observations in the data array :	1	2	3	4	5	6	7
Service time (in minutes) :	3	3.5	3.8	4	4.5	5	5.5

The median for this data would be

$$\begin{aligned}\text{Med} &= \text{value of } (n + 1)/2 \text{ th observation in the data array} \\ &= \{(7 + 1) \div 2\} \text{th} = 4 \text{th observation in the data array} = 4\end{aligned}$$

Thus, the median service time is 4 minutes per customer.

Example 3.31: Calculate the median of the following data that relates to the number of patients examined per hour in the outpatient ward (OPD) in a hospital: 10, 12, 15, 20, 13, 24, 17, 18

Solution: The data are arranged in ascending order as follows:

Observations in the data array :	1	2	3	4	5	6	7	8
Number of patients :	10	12	13	15	17	18	20	24

Since the number of observations in the data array are even, the average of $(n/2)$ th = 4th observation, i.e. 15 and $(n/2) + 1 = 5$ th observation, i.e. 17, will give the median, that is,

$$\text{Med} = (15 + 17) \div 2 = 16$$

Thus median number of patients examined per hour in OPD in a hospital are 16.

Grouped Data

To find the median value for grouped data, first identify the class interval which contains the median value or $(n/2)$ th observation of the data set. To identify such class interval, find the cumulative frequency of each class until the class for which the cumulative frequency is equal to or greater than the value of $(n/2)$ th observation. The value of the median within that class is found by using interpolation. That is, it is assumed that the observation values are evenly spaced over the entire class interval. The following formula is used to determine the median of grouped data:

$$\text{Med} = l + \frac{(n/2) - cf}{f} \times h$$

where l = lower class limit (or boundary) of the median class interval

cf = cumulative frequency of the class prior to the median class interval, that is, the sum of all the class frequencies upto, but not including, the median class interval

f = frequency of the median class

h = width of the median class interval

n = total number of observations in the distribution.

Example 3.32: A survey was conducted to determine the age (in years) of 120 automobiles. The result of such a survey is as follows:

Age of auto	:	0–4	4–8	8–12	12–16	16–20
Number of autos :		13	29	48	22	8

What is the median age for the autos?

Solution: Finding the cumulative frequencies to locate the median class as shown in Table 3.22.

Table 3.22 Calculations for Median Value

Age of Auto (in years)	Number of Autos (f)	Cumulative Frequency (cf)
0–4	13	13
4–8	29	42
8–12	48	90 ← Median class
12–16	22	112
16–20	8	120
$n = 120$		

Here the total number of observations (frequencies) are $n = 120$. Median is the size of $(n/2)$ th = $120 \div 2 = 60$ th observation in the data set. This observation lies in the class interval 8–12. Applying the formula (3-16), we have

$$\begin{aligned}\text{Med} &= l + \frac{(n/2) - cf}{n} \times h \\ &= 8 + \frac{(120 \div 2) - 42}{48} \times 4 = 8 + 1.5 = 9.5\end{aligned}$$

Example 3.33: In a factory employing 3000 persons, 5 per cent earn less than Rs 150 per day, 580 earn from Rs 151 to Rs 200 per day, 30 per cent earn from Rs 201 to Rs 250 per day, 500 earn from Rs 251 to Rs 300 per day, 20 per cent earn from Rs 301 to Rs 350 per day, and the rest earn Rs 351 or more per day. What is the median wage?

Solution: Calculation of median wage per day are shown in Table 3.23.

Table 3.23 Calculation of Median Wage

Earnings (Rs)	Percentage of Workers (Per cent)	Number of Persons (f)	Cumulative Frequency (cf)
Less than 150	5	150	150
151–200	—	580	730
201–250	30	900	(1630) ← Median class
251–300	—	500	2130
301–350	20	600	2730
351 and above	—	270	3000
$n = 3000$			

Median observation = $(n/2)\text{th} = (3000)/2 = 1500\text{th}$ observation. This observation lies in the class interval 201–250.

Now applying the formula (3-16), we have

$$\begin{aligned}\text{Med} &= l + \frac{(n/2) - cf}{f} \times h \\ &= 201 + \frac{1500 - 730}{900} \times 50 = 201 + 42.77 = \text{Rs } 243.77\end{aligned}$$

Hence, the median wage is Rs 243.77 per day.

3.9.2 Advantages, Disadvantages, and Applications of Median

Advantages

- (i) Median is unique, i.e. like mean, there is only one median for a set of data.
- (ii) The value of median is easy to understand and may be calculated from any type of data. The median in many situations can be located simply by inspection.
- (iii) The sum of the absolute differences of all observations in the data set from median value is minimum. In other words, the absolute difference of observations from the median is less than from any other value in the distribution. That is, $\sum |x - \text{Med}|$ = a minimum value.
- (iv) The extreme values in the data set does not affect the calculation of the median value and therefore it is the useful measure of central tendency when such values do occur.
- (v) The median is considered the best statistical technique for studying the qualitative attribute of an observation in the data set.
- (vi) The median value may be calculated for an open-ended distribution of data set.

Disadvantages

- (i) The median is not capable of algebraic treatment. For example, the median of two or more sets of data cannot be determined.
- (ii) The value of median is affected more by sampling variations, that is, it is affected by the number of observations rather than the values of the observations. Any observation selected at random is just as likely to exceed the median as it is to be exceeded by it.
- (iii) Since median is an average of position, therefore arranging the data in ascending or descending order of magnitude is time consuming in case of a large number of observations.
- (iv) The calculation of median in case of grouped data is based on the assumption that values of observations are evenly spaced over the entire class interval.

Applications

The median is helpful in understanding the characteristic of a data set when

- (i) observations are qualitative in nature

- (ii) extreme values are present in the data set
- (iii) a quick estimate of an average is desired.

3.10 PARTITION VALUES—QUARTILES, DECILES, AND PERCENTILES

The basic purpose of all the measures of central tendency discussed so far was to know more and more about the characteristics of a data set. Another method to analyse a data set is by arranging all the observations in either ascending or descending order of their magnitude and then dividing this ordered series into two equal parts by applying the concept of median. However, to have more knowledge about the data set, we may decompose it into more parts of equal size. The measures of central tendency which are used for dividing the data into several equal parts are called *partition values*.

In this section, we shall discuss data analysis by dividing it into *four*, *ten*, and *hundred* parts of equal size and the corresponding partition values are called *quartiles*, *deciles*, and *percentiles*. All these values can be determined in the same way as median. The only difference is in their location.

Quartiles: The values which divide an ordered data set into 4 equal parts. The 2nd quartile is the median

Quartiles The values of observations in a data set, when arranged in an ordered sequence, can be divided into four equal parts, or quarters, using three quartiles namely Q_1 , Q_2 , and Q_3 . The first quartile Q_1 divides a distribution in such a way that 25 per cent ($=n/4$) of observations have a value less than Q_1 and 75 per cent ($=3n/4$) have a value more than Q_1 , i.e. Q_1 is the median of the ordered values that are below the median.

The second quartile Q_2 has the same number of observations above and below it. It is therefore same as median value.

The quartile Q_3 divides the data set in such a way that 75 per cent of the observations have a value less than Q_3 and 25 per cent have a value more than Q_3 , i.e. Q_3 is the median of the order values that are above the median.

The generalized formula for calculating quartiles in case of grouped data is:

$$Q_i = l + \left\{ \frac{i(n/4) - cf}{f} \right\} \times h; \quad i = 1, 2, 3 \quad (3-17)$$

where cf = cumulative frequency prior to the quartile class interval

l = lower limit of the quartile class interval

f = frequency of the quartile class interval

h = width of the class interval

Deciles: The values which divide an ordered data set into 10 equal parts. The 5th decile is the median.

Deciles The values of observations in a data set when arranged in an ordered sequence can be divided into ten equal parts, using nine deciles, D_i ($i = 1, 2, \dots, 9$). The generalized formula for calculating deciles in case of grouped data is:

$$D_i = l + \left\{ \frac{i(n/10) - cf}{f} \right\} \times h; \quad i = 1, 2, \dots, 9 \quad (3-18)$$

where the symbols have their usual meaning and interpretation.

Percentiles: The values which divide an ordered data set into 100 equal parts. The 50th percentile is the median.

Percentiles The values of observations in a data when arranged in an ordered sequence can be divided into hundred equal parts using ninety nine percentiles, P_i ($i = 1, 2, \dots, 99$). In general, the i th percentile is a number that has $i\%$ of the data values at or below it and $(100 - i)\%$ of the data values at or above it. The lower quartile (Q_1), median and upper quartile (Q_3) are also the 25th percentile, 50th percentile and 75th percentile, respectively. For example, if you are told that you scored 90th percentile in a test (like the CAT), it indicates that 90% of the scores were at or below your score, while 10% were at or above your score. The generalized formula for calculating percentiles in case of grouped data is:

$$P_i = l + \left\{ \frac{i(n/100) - cf}{f} \right\} \times h; \quad i = 1, 2, \dots, 99 \quad (3-19)$$

where the symbols have their usual meaning and interpretation.

3.10.1 Graphical Method for Calculating Partition Values

The graphical method of determining various partition values can be summarized into following steps:

- (i) Draw an ogive (cumulative frequency curve) by ‘less than’ method.
- (ii) Take the values of observations or class intervals along the horizontal scale (i.e. x -axis) and cumulative frequency along vertical scale (i.e., y -axis).
- (iii) Determine the median value, that is, value of $(n/2)$ th observation, where n is the total number of observations in the data set.
- (iv) Locate this value on the y -axis and from this point draw a line parallel to the x -axis meeting the ogive at a point, say P. Draw a perpendicular on x -axis from P and it meets the x -axis at a point, say M.

The other partition values such as quartiles, deciles, and percentiles can also be obtained by drawing lines parallel to the x -axis to the distance $i(n/4)$ ($i = 1, 2, 3$); $i(n/10)$ ($i = 1, 2, \dots, 9$), and $i(n/100)$ ($i = 1, 2, \dots, 99$), respectively.

Example 3.34: The following is the distribution of weekly wages of 600 workers in a factory:

Weekly Wages (in Rs)	Number of Workers	Weekly Wages (in Rs)	Number of Workers
Below 875	69	1100 – 1175	58
875 – 950	167	1175 – 1250	24
950 – 1025	207	1250 – 1325	10
1025 – 1100	65		600

- (a) Draw an ogive for the above data and hence obtain the median value. Check it against the calculated value.
- (b) Obtain the limits of weekly wages of central 50 per cent of the workers.
- (c) Estimate graphically the percentage of workers who earned weekly wages between 950 and 1250.

[Delhi Univ., MBA, 1996]

Solution: (a) The calculations of median value are shown in Table 3.24.

Table 3.24 Calculations of Median Value

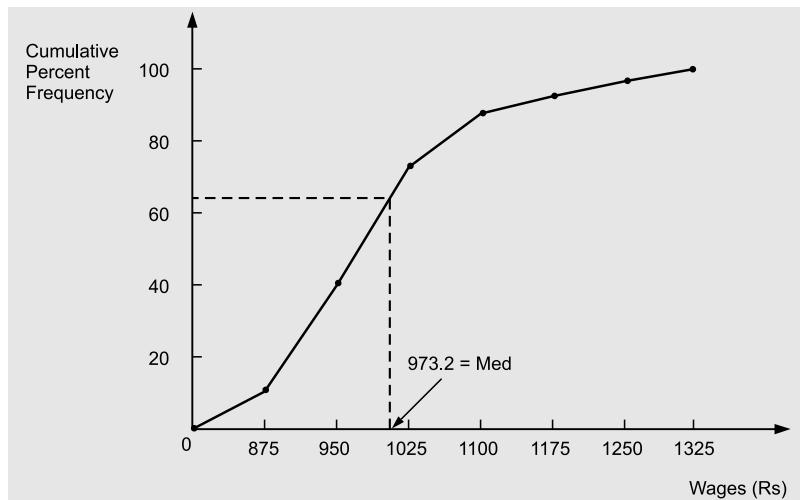
Weekly Wages (in Rs)	Number of Workers (f)	Cumulative Frequency (Less than type)	Percent Cumulative Frequency
Less than 875	69	69	11.50
Less than 950	167	236 $\leftarrow Q_1$ class	39.33
Less than 1025	207	443 \leftarrow Median class	73.83
Less than 1100	65	508 $\leftarrow Q_3$ class	84.66
Less than 1175	58	566	94.33
Less than 1250	24	590	98.33
Less than 1325	10	600	100.00

Since a median observation in the data set is the $(n/2)$ th observation = $(600 \div 2)$ th observation, that is, 300th observation. This observation lies in the class interval 950–1025. Applying the formula (3-16) to calculate median wage value, we have

$$\begin{aligned} \text{Med} &= l + \frac{(n/2) - cf}{f} \times h \\ &= 950 + \frac{300 - 236}{207} \times 75 = 950 + 23.2 = \text{Rs } 973.2 \text{ per week} \end{aligned}$$

The median wage value can also be obtained by applying the graphical method as shown in Fig. 3.1.

Fig. 3.1
Cumulative Frequency Curve



$$\begin{aligned} Q_1 &= \text{value of } (n/4)\text{th observation} \\ &= \text{value of } (600/4)\text{th} = 150\text{th observation} \end{aligned}$$

(b) The limits of weekly wages of central 50 per cent of the workers can be calculated by taking the difference of Q_1 and Q_3 . This implies that Q_1 lies in the class interval 875–950. Thus

$$\begin{aligned} Q_1 &= l + \frac{(n/4) - cf}{f} \times h \\ &= 875 + \frac{150 - 69}{167} \times 75 = 875 + 36.38 = \text{Rs } 911.38 \text{ per week} \end{aligned}$$

Similarly, $Q_3 = \text{Value of } (3n/4)\text{th observation}$
 $= \text{Value of } (3 \times 600/4)\text{th} = 450\text{th observation}$

This value of Q_3 lies in the class interval 1025–1100. Thus

$$\begin{aligned} Q_3 &= l + \frac{(3n/4) - cf}{f} \times h \\ &= 1025 + \frac{450 - 443}{65} \times 75 = 1025 + 8.08 = \text{Rs } 1033.08 \text{ per week} \end{aligned}$$

Hence the limits of weekly wages of central 50 per cent workers are Rs 911.38 and Rs 1033.08.

(c) The percentage of workers who earned weekly wages less than or equal to Rs 950 is 39.33 and who earned weekly wages less than or equal to Rs 1250 is 98.33. Thus the percentage of workers who earned weekly wages between Rs 950 and Rs 1250 is $(98.33 - 39.33) = 59$.

Example 3.35: You are working for the transport manager of a ‘call centre’ which hires cars for the staff. You are interested in the weekly distances covered by these cars. Kilometers recorded for a sample of hired cars during a given week yielded the following data:

Kilometers Covered	Number of Cars	Kilometers Covered	Number of Cars
100–110	4	150–160	8
110–120	0	160–170	5
120–130	3	170–180	0
130–140	7	180–190	2
140–150	11		40

- Form a cumulative frequency distribution and draw a cumulative frequency ogive.
- Estimate graphically the number of cars which covered less than 165 km in the week.
- Calculate Q_1 , Q_2 , Q_3 and P_{75} .

Solution: (a) The calculations to obtain a cumulative frequency distribution and to draw give are shown in table 3.25.

Table 3.25

Kilometers Covered Less than	Number of Cars	Cumulative Frequency	Percent Cumulative Frequency
110	4	4	10.0
120	0	4	10.0
130	3	7	17.5
140	7	14 $\leftarrow Q_1$	35.0
150	11	25 $\leftarrow Me = Q_2$	62.5
160	8	33 $\leftarrow Q_3$	82.5 $\leftarrow P_{75}$
170	5	38	95.0
180	0	38	95.0
190	2	40	100.0

Plotting cumulative frequency values on the graph paper, frequency polygon is as shown in Fig. 3.2.

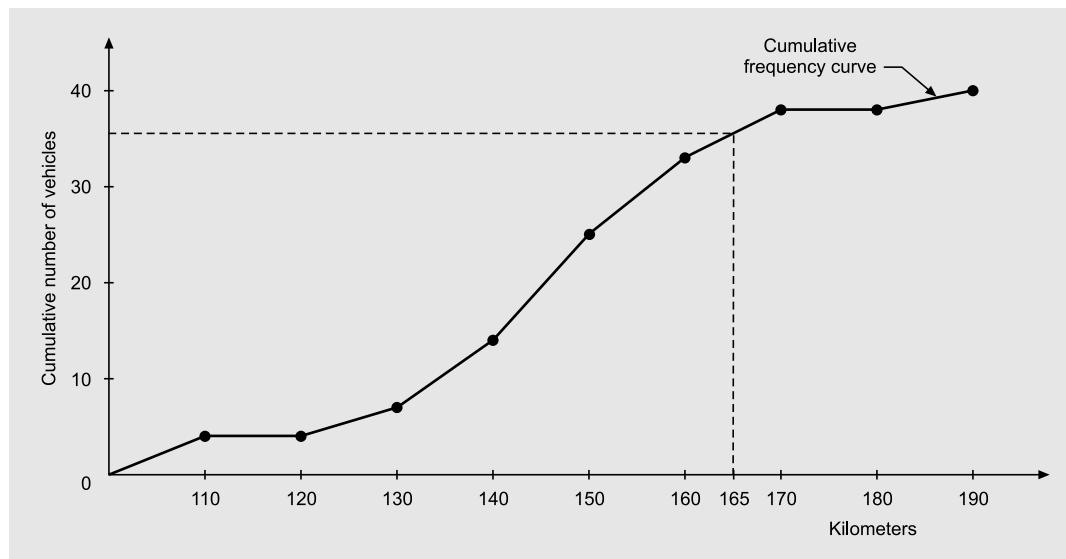


Fig. 3.2
Cumulative Frequency Curve

(b) The number of cars which covered less than 165 km in the week are 35 as shown in the Fig. 3.2.

(c) Since there are 40 observations in the data set, we can take 10th, 20th and 30th cumulative values to corresponds to Q_1 , Q_2 and Q_3 respectively. These values from the graph gives $Q_1 = 134$, $Q_2 = 146$ and $Q_3 = 156$.

$$P_{75} = \frac{(75n/100) - cf}{f} \times h = 150 + \frac{30 - 25}{8} \times 10 = 156.25$$

This implies that 75 per cent of cars covered less than or equal to 156.25 kilometers.

Example 3.36: The following distribution gives the pattern of overtime work per week done by 100 employees of a company. Calculate median, first quartile, and seventh decile.

Overtime hours : 10–15 15–20 20–25 25–30 30–35 35–40

No. of employees : 11 20 35 20 8 6

[Kurukshetra Univ., MBA, 1997]

Calculate Q_1 , D_7 and P_{60} .

Solution: The calculations of median, first quartile (Q_1), and seventh decile (D_7) are shown in Table 3.26.

Table 3.26

Overtime Hours	Number of Employees	Cumulative Frequency (Less than type)
10–15	11	11
15–20	20	31 ← Q_1 class
20–25	35	66 ← Median and P_{60} class
25–30	20	86 ← D_7 class
30–35	8	94
35–40	6	100
	100	

Since the number of observations in this data set are 100, the median value is $(n/2)$ th = $(100 \div 2)$ th = 50th observation. This observation lies in the class interval 20–25. Applying the formula (3-16) to get median overtime hours value, we have

$$\begin{aligned} \text{Med} &= l + \frac{(n/2) - cf}{f} \times h \\ &= 20 + \frac{50 - 31}{35} \times 5 = 20 + 2.714 = 22.714 \text{ hours} \end{aligned}$$

Q_1 = value of $(n/4)$ th observation = value of $(100/4)$ th = 25th observation

$$\text{Thus } Q_1 = l + \frac{(n/4) - cf}{f} \times h = 15 + \frac{25 - 11}{20} \times 5 = 15 + 3.5 = 18.5 \text{ hours}$$

D_7 = value of $(7n/10)$ th observation = value of $(7 \times 100)/10$ = 70th observation

$$\text{Thus } D_7 = l + \frac{(7n/10) - cf}{f} \times h = 25 + \frac{70 - 66}{20} \times 5 = 25 + 1 = 26 \text{ hours}$$

P_{60} = Value of $(60n/100)$ th observation = $60 \times (100/100) = 60$ th observation

$$\text{Thus } P_{60} = l + \frac{(60 \times n/100) - cf}{f} \times h = 20 + \frac{60 - 31}{35} \times 5 = 24.14 \text{ hours}$$

Conceptual Questions 3C

16. Define median and discuss its advantages and disadvantages.
17. Why is it necessary to interpolate in order to find the median of a grouped data?
18. When is the use of median considered more appropriate than mean?
19. Write a short criticism of the following statement: 'Median is more representative than mean because it is relatively less affected by extreme values'.
20. What are quartiles of a distribution? Explain their uses.
21. It has been said that the same percentage of frequencies falls between the first and ninth decile for symmetric and skewed distributions. Criticize or explain this statement. Generalize your answer to other percentiles.
22. Describe the similarities and differences among median, quartiles, and percentiles as descriptive measures of position.
23. You obtained the following answers to a statement while conducting a survey on reservation for women in politics strongly disagree, disagree, mildly disagree, agree some what, agree, strongly agree. Of these answers, which is the median?

Self-Practice Problems 3D

- 3.29** On a university campus 200 teachers are asked to express their views on how they feel about the performance of their Union's president. The views are classified into the following categories:

Disapprove strongly	= 94
Disapprove	= 52
Approve	= 43
Approve strongly	= 11

What is the median view?

- 3.30** The following are the profit figures earned by 50 companies in the country

Profit (in Rs lakh)	Number of Companies
10 or less	4
20 or less	10
30 or less	30
40 or less	40
50 or less	47
60 or less	50

Calculate

- (a) the median, and
- (b) the range of profit earned by the middle 80 per cent of the companies. Also verify your results by graphical method.

- 3.31** A number of particular items has been classified according to their weights. After drying for two weeks the same items have again been weighted and similarly classified. It is known that the median weight in the first weighing was 20.83 g, while in the second weighing it was 17.35 g. Some frequencies, a and b , in the first weighing and x and y in the second weighing are missing. It is known that $a = x/3$ and $b = y/2$. Find out the values of the missing frequencies.

Class	Frequencies		Class	Frequencies	
	I	II		I	II
0–5	a	x	15–20	52	50
5–10	b	y	20–25	75	30
10–15	11	40	25–30	22	28

- 3.32** The length of time taken by each of 18 workers to complete a specific job was observed to be the following:

Time (in min) : 5–9 10–14 15–19 20–24 25–29
Number of workers : 3 8 4 2 1

- (a) Calculate the median time
- (b) Calculate Q_1 and Q_3

- 3.33** The distribution of the insurance money paid by an automobile insurance company to owners of automobiles in a particular year is given below:

Amount Paid (in Rs)	Frequency	Amount Paid (in Rs)	Frequency
below 1500	52	3500–3499	816
1500–1999	108	4000–4499	993
2000–2499	230	4500–4999	825
2500–2999	528	5000 and above	650
3000–3499	663		

Calculate the median amount of money paid.

- 3.34** The following distribution is with regard to weight (in g) of mangoes of a given variety. If mangoes less than 443 g in weight be considered unsuitable for the foreign market, what is the percentage of total yield suitable for it? Assume the given frequency distribution to be typical of the variety.

Weight (in g)	Number of Mangoes	Weight (in g)	Number of Mangoes
410–419	10	450–459	45
420–429	20	460–469	18
430–439	42	470–479	7
440–449	54		

Draw an ogive of ‘more than’ type of the above data and deduce how many mangoes will be more than 443 g.

- 3.35** Gupta Machine Company has a contract with his customers to supply machined pump gears. One requirement is that the diameter of gears be within specific limits. Here are the diameters (in inches) of a sample of 20 gears:

4.01	4.00	4.02	4.02	4.03	4.00
3.98	3.99	3.99	4.01	3.99	3.98
3.97	4.00	4.02	4.01	4.02	4.00
4.01	3.99				

What can Gupta say to his customers about the diameters of 95 per cent of the gears they are receiving?

[Delhi Univ., MBA, 1998]

- 3.36** Given the following frequency distribution with some missing frequencies:

Class	Frequency	Class	Frequency
10–20	185	50–60	136
20–30	—	60–70	—
30–40	34	70–80	50
40–50	180		

If the total frequency is 685 and median is 42.6, find out the missing frequencies.

Hints and Answers

3.29 Disapprove

3.30 Med = 27.5; $P_{90} - P_{10} = 47.14 - 11.67 = 35.47$

3.31 $a = 3, b = 6; x = 9, y = 12$

3.32 (a) 13.25 (b) $Q_3 = 17.6, Q_1 = 10.4$

3.34 52.25%; 103

3.35 Diameter size : 3.97 3.98 3.99 4.00 4.01 4.02 4.03

Frequency : 1 2 4 4 4 4 1

$$\bar{x} = \frac{\sum x}{n} = \frac{80.04}{20} = 4.002 \text{ inches;}$$

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

$$= \sqrt{\frac{1}{19} \{ (320.32 - 20(4.002)^2) \}}$$

= 0.016 inches 95% of the gears will have diameter in the interval : $\bar{x} \pm 2s = (4.002 \pm 0.016)$

3.36 20–30(77)

3.11 MODE

Mode value: A measure of location recognised by the location of the most frequently occurring value of a set of data.

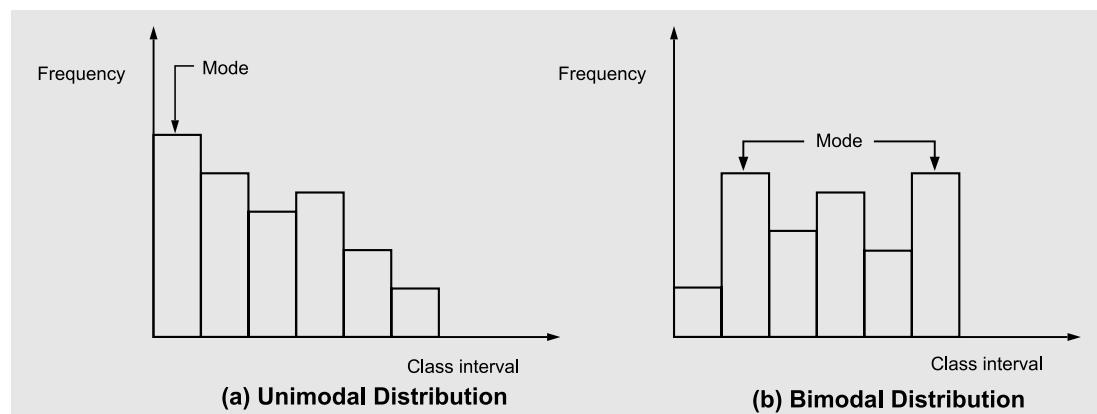
The **mode** is that value of an observation which occurs most frequently in the data set, that is, the point (or class mark) with the highest frequency.

The concept of mode is of great use to large scale manufacturers of consumable items such as ready-made garments, shoe-makers, and so on. In all such cases it is important to know the size that fits most persons rather than ‘mean’ size.

There are many practical situations in which arithmetic mean does not always provide an accurate characteristic (reflection) of the data due to the presence of extreme values. For example, in all such statements like ‘average man prefers ... brand of cigarettes’, ‘average production of an item in a month’, or ‘average service time at the service counter’. The term ‘average’ means majority (i.e. mode value) and not the arithmetic mean. Similarly, the median may not represent the characteristics of the data set completely owing to an uneven distribution of the values of observations. For example, suppose in a distribution the values in the lower half vary from 10 to 100 (say), while the same number of observations in the upper half vary from 100 to 7000 (say) with most of them close to the higher limit. In such a distribution, the median value of 100 will not provide an indication of the true nature of the data. Such shortcomings stated above for mean and median are removed by the use of *mode*, the third measure of central tendency.

The mode is a poor measure of central tendency when most frequently occurring values of an observation do not appear close to the centre of the data. The mode need not even be a unique value. Consider the frequency distributions shown in Fig. 3.3(a) and (b). The distribution in Fig. 3.3(a) has its mode at the lowest class and certainly cannot be considered representative of central location. The distribution shown in Fig. 3.3(b) has two modes. Obviously neither of these values appear to be representative of the central location of the data. For these reasons the mode has limited use as a measure of central tendency for decision-making. However, for descriptive analysis, mode is a useful measure of central tendency.

Fig. 3.3
Frequency Distribution



Calculation of Mode It is always preferable to calculate mode from grouped data. Table 3.27, for example, shows the sales per day of an item for 20 days period. The mode of this data is 71 since this value occurs more frequently (four times than any other value). However, it fails to reveal the fact that most of the values are under 70.

Table 3.27 Sales During 20 Days Period
(Data arranged in ascending order)

53,	56,	57,	58,	58,	60,	61,	63,	63,	64
64,	65,	65,	67,	68,	71,	71,	71,	71,	74

Converting this data into a frequency distribution as shown in Table 3.28:

Table 3.28 Frequency Distribution of Sales Per Day

<i>Sales volume (Class interval)</i>	: 53–56	57–60	61–64	65–68	69–72	72 and above
<i>Number of days (Frequency)</i>	: 2	4	5	4	4	1

Table 3.28 shows that a sale of 61–64 units of the item was achieved on 5 days. Thus this class is more representative of the sales per day.

In the case of grouped data, the following formula is used for calculating mode:

$$\text{Mode} = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h$$

where l = lower limit of the model class interval

f_{m-1} = frequency of the class preceding the mode class interval

f_{m+1} = frequency of the class following the mode class interval

h = width of the mode class interval

Example 3.37: Using the data of Table 3.28, calculate the mode of sales distribution of the units of item during the 20 days period.

Solution: Since the largest frequency corresponds to the class interval 61–64, therefore it is the mode class. Then we have, $l = 61$, $f_m = 5$, $f_{m-1} = 4$, $f_{m+1} = 4$ and $h = 3$. Thus

$$\begin{aligned} M_0 &= l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h \\ &= 61 + \frac{5 - 4}{10 - 4 - 4} \times 3 = 61 + 1.5 = 62.5 \end{aligned}$$

Hence, the modal sale is of 62.5 units.

Example 3.38: In 500 small-scale industrial units, the return on investment ranged from 0 to 30 per cent; no unit sustaining loss. Five per cent of the units had returns ranging from zero per cent to (and including) 5 per cent, and 15 per cent of the units earned returns exceeding 5 per cent but not exceeding 10 per cent. The median rate of return was 15 per cent and the upper quartile 2 per cent. The uppermost layer of returns exceeding 25 per cent was earned by 50 units.

(a) Present the information in the form of a frequency table as follows:

Exceeding 0 per cent but not exceeding 5 per cent

Exceeding 5 per cent but not exceeding 10 per cent

Exceeding 10 per cent but not exceeding 15 per cent

and so on.

(b) Find the rate of return around which there is maximum concentration of units.

Solution: (a) The given information is summarized in the form of a frequency distribution as shown in Table 3.29.

Table 3.29

<i>Rate of Return</i>	<i>Industrial Units</i>
Exceeding 0 per cent but not exceeding 5 per cent	$500 \times \frac{5}{100} = 25$
Exceeding 5 per cent but not exceeding 10 per cent	$500 \times \frac{15}{100} = 75$
Exceeding 10 per cent but not exceeding 15 per cent	$250 - 100 = 150$
Exceeding 15 per cent but not exceeding 20 per cent	$375 - 250 = 125$
Exceeding 20 per cent but not exceeding 25 per cent	$500 - 375 - 50 = 75$
Exceeding 25 per cent but not exceeding 30 per cent	50

(b) Calculating mode to find out the rate of return around which there is maximum concentration of the units. The mode lies in the class interval 10–15. Thus

$$\begin{aligned} M_o &= l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \\ &= 10 + \frac{150 - 75}{2 \times 150 - 75 - 125} \times 5 = 10 + 3.75 = 13.75 \text{ per cent} \end{aligned}$$

3.11.1 Graphical Method for Calculating Mode Value

The procedure of calculating mode using the graphical method is summarized below:

- (i) Draw a histogram of the data, the tallest rectangle will represent the modal class.
- (ii) Draw two diagonal lines from the top right corner and left corner of the tallest rectangle to the top right corner and left corner of the adjacent rectangles.
- (iii) Draw a perpendicular line from the point of intersection of the two diagonal lines on the x -axis. The value on the x -axis marked by the line will represent the modal value.

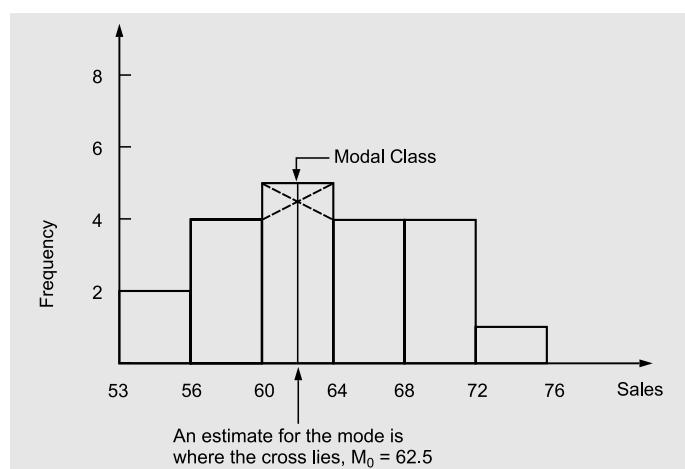
Example 3.39: Calculate the mode using the graphical method for the following distribution of data:

Sales (in units) :	53–56	57–60	61–64	65–68	69–72	73–76
Number of days :	2	4	5	4	4	1

Solution: Construct a histogram of the data shown in Fig. 3.4 and draw other lines for the calculation of mode value.

The mode value from Fig. 3.4 is 62.5 which is same as calculated in Example 3.37.

Figure 3.4
Graph for Modal Value



3.11.2 Advantages and Disadvantages of Mode Value

Advantages

- (i) Mode value is easy to understand and to calculate. Mode class can also be located by inspection.
- (ii) The mode is not affected by the extreme values in the distribution. The mode value can also be calculated for open-ended frequency distributions.
- (iii) The mode can be used to describe quantitative as well as qualitative data. For example, its value is used for comparing consumer preferences for various types of products, say cigarettes, soaps, toothpastes, or other products.

Disadvantages

- (i) Mode is not a rigidly defined measure as there are several methods for calculating its value.
- (ii) It is difficult to locate modal class in the case of multi-modal frequency distributions.
- (iii) Mode is not suitable for algebraic manipulations.
- (iv) When data sets contain more than one modes, such values are difficult to interpret and compare.

3.12 RELATIONSHIP BETWEEN MEAN, MEDIAN, AND MODE

In a *unimodal* and symmetrical distribution, the values of mean, median, and mode are equal as indicated in Fig. 3.5. In other words, when all these three values are not equal to each other, the distribution is not symmetrical.

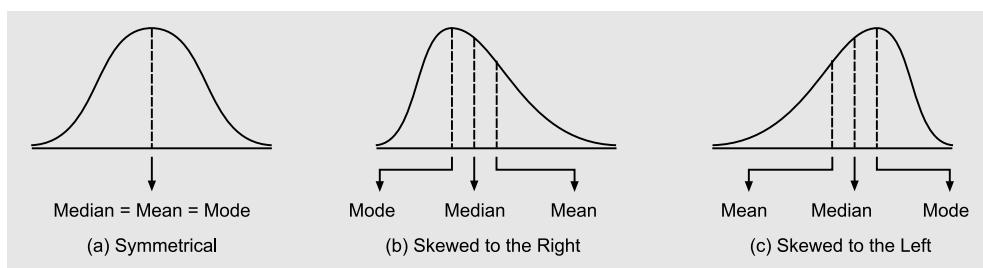


Figure 3.5
A comparison of Mean, Median, and Mode for three Distributional Shapes

A distribution that is not symmetrical, but rather has most of its values either to the right or to the left of the mode, is said to be *skewed*. For such asymmetrical distribution, Karl Pearson has suggested a relationship between these three measures of central tendency as:

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median}) \quad (3-22)$$

or

$$\text{Mode} = 3\text{Median} - 2\text{Mean}$$

This implies that the value of any of these three measures can be calculated provided we know any two values out of three. The relationship (3-22) is shown in Fig. 3.5(b) and (c).

If most of the values of observations in a distribution fall to the right of the mode as shown in Fig. 3.5(b), then it is said to be skewed to the right or *positively skewed*. Distributions that are skewed right contain a few unusually large values of observations. In this case, mode remains under the peak (i.e., representing highest frequency) but the median (value that depends on the number of observations) and mean move to the right (value that is affected by extreme values). The order of magnitude of these measures will be

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

But if the distribution is skewed to the left or *negatively skewed* (i.e., values of lower magnitude are concentrated more to the left of the mode) then mode is again under the peak whereas median and mean move to the left of mode. The order of magnitude of these measures will be

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

In both these cases, the difference between mean and mode is 3 times the difference between mean and median.

In general, for a single-peaked skewed distribution (non-symmetrical), the median is preferred to the mean for measuring location because it is neither influenced by the frequency of occurrence of a single observation value as mode nor it is affected by extreme values.

3.13 COMPARISON BETWEEN MEASURES OF CENTRAL TENDENCY

In this chapter, we have already presented three methods to understand the characteristics of a data set. However, the choice of which method to use for describing a distribution of values of observations in a data set is not always easy. The choice to use any one of these three is mainly guided by their characteristics. The characteristics of these three differ from each other with regard to three factors:

- (i) Presence of outlier data values
- (ii) Shape of the frequency distribution of data values
- (iii) Status of theoretical development

Outlier: A very small or very large value in the data set.

1. **The Presence of Outlier Data Values:** The data values that differ in a big way from the other values in a data set are known as *outliers* (either very small or very high values). As mentioned earlier, the median is not sensitive to outlier values because its value depend only on the number of observations and the value always lies in the middle of the ordered set of values, whereas mean, which is calculated using all data values is sensitive to the outlier values in a data set. Obviously, smaller the number of observations in a data set, greater the influence of any outliers on the mean. The median is said to be *resistant* to the presence of outlier data values, but the mean is not.
2. **Shape of Frequency Distribution:** The effect of the shape of frequency distribution on mean, median, and mode is shown in Fig. 3.5. In general, the median is preferred to the mean as a way of measuring location for single peaked, skewed distributions. One of the reasons is that it satisfies the criterion that the *sum of absolute* difference (i.e., absolute error of judgment) of median from values in the data set is minimum, that is, $\Sigma |x - \text{Med}| = \min$. In other words, the smallest sum of the absolute errors is associated with the median value is the data set as compared to either mean or mode. When data is multi-modal, there is no single measure of central location and the mode can vary dramatically from one sample to another, particularly when dealing with small samples.
3. **The Status of Theoretical Development:** Although the three measures of central tendency—Mean, Median, and Mode, satisfy different mathematical criteria but the objective of any statistical analysis in *inferential statistics* is always to minimize the *sum of squared deviations* (*errors*) taken from these measures to every value in the data set. The criterion of the sum of squared deviations is also called *least squares criterion*. Since A.M. satisfies the least squares criterion, it is mathematically consistent with several techniques of statistical inference.

As with the median, it can not be used to develop theoretical concepts and models and so is only used for basic descriptive purposes.

Conceptual Questions 3C

24. Give a brief description of the different measures of central tendency. Why is arithmetic mean so popular?
25. How would you explain the choice of arithmetic mean as the best measure of central tendency. Under what circumstances would you deem fit the use of median or mode?
26. What are the advantages and disadvantages of the three common averages: Mean, Median, and Mode?

27. Describe the relationship between the mean and median of a set of data to indicate the skewness of the distribution of values.
28. Identify the mathematical criteria associated with mean, median, and mode and briefly explain the meaning of each criterion.
29. It is said that the use of a particular average depends upon the particular problem in hand. Comment and indicate at least one instance of the use of mean, median, mode, geometric, and harmonic mean.
30. How would you account for the predominant choice of arithmetic mean as a measure of central tendency? Under what circumstances would it be appropriate to use mode or median? [Delhi Univ., MBA, 2000]
31. Under what circumstances would it be appropriate to use mean, median, or mode? Discuss [Delhi Univ., MBA, 1996]
32. Explain the properties of a good average. In the light of these properties which average do you think is best and why? [Jodhpur Univ., MBA, 1996]
33. Give a brief note of the measures of central tendency together with their merits and demerits. Which is the best

measure of central tendency and why?

[Osmania Univ., MBA, 1998]

34. What is a statistical average? What are the desirable properties for an average to possess? Mention the different types of averages and state why arithmetic mean is most commonly used amongst them.
35. What is the relationship between mean, median, and mode? Under what circumstances are they equal?
36. It has been said that the less the variability, the more an average is representative of a set of data. Comment on the meaning of this statement.
37. Which measure of central tendency is usually preferred if the distribution is known to be single peaked and skewed? Why?
38. Suppose the average amount of cash (in pocket, wallet, purse, etc.) possessed by 60 students attending a class is Rs 125. The median amount carried is Rs 90.
- What characteristics of the distribution of cash carried by the students can be explained. Why is mean larger than the median?
 - Identify the process or population to which inferences based on these results might apply.

Self-Practice Problems 3E

- 3.37** Given below is the distribution of profits (in '000 rupees) earned by 94 per cent of the retail grocery shops in a city.

Profits	Number of Shops	Profit	Number of Shops
0–10	0	50–60	68
10–20	5	60–70	83
20–30	14	70–80	91
30–40	27	80–90	94
40–50	48		

Calculate the modal value.

- 3.38** Compute mode value from the following data relating to dividend paid by companies in a particular financial year.

Dividend (in per cent) Value	Number of Companies (in per cent)	Dividend of the Share Value	Number of Companies
5.0–7.5	182	15.0–17.5	280
7.5–10.0	75	17.5–20.0	236
10.0–12.5	59	20.0–22.5	378
12.5–15.0	127	22.5–25.0	331

- 3.39** Following is the cumulative frequency distribution of the preferred length of kitchen slabs obtained from the preference study on 50 housewives:

Length (metres) more than	Number of Housewives
1.0	50
1.5	46
2.0	40
2.5	42
3.0	10
3.5	3

A manufacturer has to take a decision on what length of slabs to manufacture. What length would you recommend and why?

- 3.40** The management of Doordarshan holds a preview of a new programme and asks viewers for their reaction. The following results by age groups, were obtained.

Age group	Under 20	20–39	40–59	60 and above
Liked the program :	140	75	50	40
Disliked the program :	60	50	50	20

Using a suitable measure of central tendency, suggest that towards which age group the management should aim its advertising campaign.

- 3.41** A sample of 100 households in a given city revealed the following number of persons per household:

Number of Persons	No. of Households
1	16
2	28
3–4	37
5–6	12
7–11	7

- (a) What is the modal category for the 100 households observed?
- (b) What proportion of the households have more than four persons.
- 3.42** The number of solar heating systems available to the public is quite large, and their heat storage capacities are quite varied. Here is a distribution of heat storage capacity (in days) of 28 systems that were tested recently by a testing agency

Days	Frequency	Days	Frequency
0–0.99	2	4–4.99	5
1–1.99	4	5–5.99	3
2–2.99	6	6–6.99	1
3–3.99	7		

The agency knows that its report on the tests will be widely circulated and used as the basis for solar heat allowances.

- (a) Compute the mean, median, and mode of these data.
- (b) Select the answer from part (a) which best reflects the central tendency of the test data and justify your choice.
- 3.43** Mr Pandey does statistical analysis for an automobile racing team. The data on fuel consumption (in km per litre) for the team's cars in recent races are as follows:

14.77 16.11 16.11 15.05 15.99 14.91
15.27 16.01 15.75 14.89 16.05 15.22
16.02 15.24 16.11 15.02

- (a) Calculate the mean and median fuel consumption.
- (b) Group the data into five equally-sized classes. What is the fuel consumption value of the modal class?
- (c) Which of the three measures of central tendency is best to use? Explain.

- 3.44** An agriculture farm sells grab bags of flower bulbs. The bags are sold by weight; thus the number of bulbs in each bag can vary depending on the varieties included. Below are the number of bulbs in each of the 20 bags sampled:

21 33 37 56 47 25 33 32 47 34
36 23 26 33 37 26 37 37 43 45

- (a) What are the mean and median number of bulbs per bag?
- (b) Based on your answer, what can you conclude about the shape of the distribution of number of bulbs per bag?
- 3.45** The table below is the frequency distribution of ages to the nearest birthday for a random sample of 50 employees in a large company

Age to nearest birthday	:	20–29	30–39	40–49	50–59	60–69
Number of employees	:	5	12	13	8	12

- Compute the mean, median, and mode for these data.
- 3.46** A track coach is in the process of selecting one of the two sprinters for the 200 meter race at the upcoming games. He has the following data of the results of five races (time in seconds) of the two sprinters run with 15 minutes rest intervals in between.

Athlete	Races				
	1	2	3	4	5
Vibhor	24.2	24.1	24.1	28.9	24.2
Prasant	24.4	24.5	24.5	24.6	24.5

Based on these data, which of the two sprinters should the coach select? Why?

- 3.47** The following data are the yields (in per cent) in the money market of 10 companies listed at the Bombay Stock Exchange (BSE) as on 18 October 2001, the day before the BSE index average passed the 3000 mark.

Company	Money Market Yield (Per cent)
Tata Power	10.0
HCL Infosys	7.5
ITC	5.7
NIIT	5.4
Cipla	4.6
Reliance Petro	4.1
Reliance	4.0
Dr. Reddy's Lab	3.9
Digital Glob	3.0
ICICI	2.9

- (a) Compute the mean, median, mode, quartile, and decile deviation for these yields.
- (b) What other information would you want to know if you were deciding to buy shares of one of these companies? Prepare a list of questions that you would like to ask a broker.

- 3.48** In order to estimate how much water will need to be supplied to a locality in East Delhi area during the summer of 2002, the minister asked the General Manager of the water supply department to find out how much water a sample of families currently uses. The sample of 20 families used the following number of gallons (in thousands) in the past years.

9.3 19.6 14.5 17.8 14.7 15.0 13.9 12.7
10.0 13.0 25.0 16.3 11.2 20.2 15.4 11.6
16.5 11.0 12.2 10.9

- (a) What is the mean and median amount of water used per family?
- (b) Suppose that 10 years from now, the government expects that there will be 1800 families living in that colony. How many gallons of water will be needed annually, if rate of consumption per family remains the same?

- (c) In what ways would the information provided in (a) and (b) be useful to the government? Discuss.

- (d) Why might the government have used the data from a survey rather than just measuring the total consumption in Delhi?

Hints and Answers

3.37 $M_0 = 3 \text{ Med} - 2\bar{x} = \text{Rs } 50.04 \text{ thousand}$

3.38 $M_0 = 21.87$ (per cent of the share value)

- 3.41** (a) Persons between 3–4
 (b) 19 per cent house holds

Formulae Used

1. Summation of n numbers

$$\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$$

Simplified expression for the summation of n numbers

$$\sum x_i = x_1 + x_2 + \dots + x_n$$

2. Sample mean, $\bar{x} = \frac{\sum x_i}{n}$

Population mean, $\mu = \frac{\sum x_i}{N}$

Sample mean for grouped data, $\bar{x} = \frac{\sum f_i m_i}{n}$

where $n = \sum f_i$ and m_i = mid-value of class intervals

3. Weighted mean for a population or a sample,

$$\bar{x}_w \text{ or } \mu_w = \frac{\sum w_i x_i}{\sum w_i}$$

where w_i = weight for observation i

4. Position of the median in an ordered set of observation belong to a population or a sample is, $\text{Med} = x_{(n/2) + (1/2)}$

Median for grouped data, $\text{Med} = l + \left[\frac{(n/2) - cf}{f} \right] h$

5. Quartile for a grouped data

$$Q_i = l + \left[\frac{i(n/4) - cf}{f} \right] h ; \quad i = 1, 2, 3$$

Decile for a grouped data

$$D_i = l + \left[\frac{i(n/10) - cf}{f} \right] h ; \quad i = 1, 2, \dots, 9$$

Percentile for a grouped data

$$P_i = l + \left[\frac{i(n/100) - cf}{f} \right] h ; \quad i = 1, 2, \dots, 99$$

6. Mode for a grouped data

$$M_0 = l + \left[\frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \right] h$$

Mode for a multimode frequency distribution

$$M_0 = 3 \text{ Median} - 2 \text{ Mean}$$

Review Self-Practice Problems

- 3.49** The following is the data on profit margin (in per cent) of three products and their corresponding sales (in Rs) during a particular period.

Product	Profit Margin (Per cent)	Sales (Rs in thousand)
A	12.5	2,000
B	10.3	6,000
C	6.4	10,000

- (a) Determine the mean profit margin.
 (b) Determine the weighted mean considering the rupee sales as weight for each product.
 (c) Which of the means calculated in part (a) and (b) is the correct one?

- 3.50** The number of cars sold by each of the 10 car dealers during a particular month, arranged in ascending order, is 12, 14, 17, 20, 20, 20, 22, 22, 24, 25. Considering this scale to be the statistical population of interest, determine the mean, median, and mode for the number of cars sold.

- (a) Which value calculated above best describes the 'typical' sales volume per dealer?
 (b) For the given data, determine the values at the (i) quartile Q_1 and (ii) percentile P_{30} for these sales amounts.

- 3.51** A quality control inspector tested nine samples of each of three designs A, B and C of certain bearing for a new electrical winch. The following data are the number of hours it took for each bearing to fail when the winch

motor was run continuously at maximum output, with a load on the winch equivalent to 1.9 times the intended capacity.

A :	16	16	53	15	31	17	14	30	20
B :	18	27	23	21	22	26	39	17	28
C :	31	16	42	20	18	17	16	15	19

Calculate the mean and median for each group and suggest which design is best and why?

[IIPM, PGDM, 2002]

- 3.52** Calculate the mean, median, and mode for the following data pertaining to marks in statistics. There are 80 students in a class and the test is of 140 marks.

Marks more than :	0	20	40	60	80	100	120
Number of students :	80	76	50	28	18	9	3

[M.D. Univ., MBA, 1994]

- 3.53** A company invests one lakh rupees at 10 per cent annual rate of interest. What will be the total amount after 6 years if the principal is not withdrawn?

- 3.54** Draw an ogive for the following distribution. Read the median from the graph and verify your result by the mathematical formula. Also obtain the limits of income of the central 50% of employees.

Weekly Income (Rs)	Number of Employees	Weekly Income (Rs)	Number of Employees
Below 550	6	700–750	16
550–600	10	750–800	12
600–650	22	800 and above	15
650–700	30		

[Delhi Univ., MBA, 1999]

- 3.55** In the production of light bulbs, many bulbs are broken. A production manager is testing a new type of conveyor system in the hope of reducing the percentage of bulbs broken each day. For ten days he observes bulb breakage with the current conveyor. He then records bulb breakage for ten days with the new system, after allowing a few days for the operator to learn to use it. His data are as follows:

Conveyor System	Percentage of Bulbs Broken Daily									
Old	8.7	11.1	4.4	3.7	9.2	6.6	7.8	4.9	6.9	8.3
New	10.8	6.2	3.2	4.6	5.3	6.5	4.6	7.1	4.9	7.2

- (a) Compute the mean and median for each conveyor system.
 (b) Based on these results, do you think this test establishes that the new system lowers the breakage rate? Explain.
- 3.56** The following are the weekly wages in rupees of 30 workers of a firm:

140	139	126	114	100	88	62	77	99
103	108	129	144	148	134	63	69	148
132	118	142	116	123	104	95	80	85
106	123	133						

The firm gave bonus of Rs 10, 15, 20, 25, 30, and 35 for individuals in the respective salary slabs: exceeding 60 but not exceeding 75; exceeding 75 but not exceeding 90; and so on up to exceeding 135 and not exceeding 150. Find the average bonus paid.

- 3.57** The mean monthly salaries paid to 100 employees of a company was Rs 5,000. The mean monthly salaries paid to male and female employees were Rs 5,200 and Rs 4,200 respectively. Determine the percentage of males and females employed by the company.
- 3.58** The following is the age distribution of 2,000 persons working in a large textile mill:

Age Group	No. of Persons	Age Group	No. of Persons
15 but less than 20	80	45 but less than 50	268
20 but less than 25	250	50 but less than 55	150
25 but less than 30	300	55 but less than 60	75
30 but less than 35	325	60 but less than 65	25
35 but less than 40	287	65 but less than 70	20
40 but less than 45	220		

Because of heavy losses the management decides to bring down the strength to 40 per cent of the present number according to the following scheme:

- (i) To retrench the first 10 per cent from lowest age group 15–20.
- (ii) To absorb the next 40 per cent in other branches.
- (iii) To make 10 per cent from the highest age group, 40–45 retire prematurely.

What will be the age limits of persons retained in the mill and of those transferred to other branches? Also calculate the average age of those retained.

- 3.59** A factory pays workers on piece rate basis and also a bonus to each worker on the basis of individual output in each quarter. The rate of bonus payable is as follows:

Output (in units)	Bonus (Rs)	Output (in units)	Bonus (Rs)
70–74	40	90–94	70
75–79	45	95–99	80
80–84	50	100–104	100
85–89	60		

The individual output of a batch of 50 workers is given below:

94	83	78	76	88	86	93	80	91	82
89	97	92	84	92	80	85	83	98	103
87	88	88	81	95	86	99	81	87	90
84	97	80	75	93	101	82	82	89	72
85	83	75	72	83	98	77	87	71	80

By suitable classification you are required to find:

- (a) Average bonus per worker for the quarter
- (b) Average output per worker.

[Pune Univ., MBA, 1998]

3.60 An economy grows at the rate of 2 per cent in the first year, 2.5 per cent in the second year, 3 per cent in the third year, 4 per cent in the fourth year . . . and 10 per cent in the tenth year. What is the average rate of growth of the company?

3.61 A man travelled by car for 3 days. He covered 480 km each day. On the first day he drove for 10 hours at 48 km an hour, on the second day he drove for 12 hours at 40 km an hour, and on the last day he drove for 15 hours at 32 km per hour. What was his average speed? [Bangalore Univ., BCom, 1996]

3.62 The monthly income of employees in an industrial concern are given below. The total income of 10 employees in the class over Rs 25,000 is Rs 3,00,000. Compute the mean income. Every employee belonging to the top 25 per cent of the earners is required to pay 5 per cent of his income to the workers' relief fund. Estimate the contribution to this fund.

Income (Rs)	Frequency	Income (Rs)	Frequency
Below 5000	90	15,000–20,000	80
5000–10000	150	20,000–25,000	70
10000–15000	100	25,000 and above	10

[Kakatiya Univ., MCom, 1997]

3.63 In a factory there are 100 skilled, 250 semi-skilled, and 150 unskilled workers. It has been observed that on an average a unit length of a particular fabric is woven by a skilled worker in 3 hours, by a semi-skilled worker in

4 hours, and by an unskilled worker in 5 hours. Unskilled workers are expected to become semi-skilled workers and semi-skilled workers are expected to become skilled. How much less time will be required after 2 years of training for weaving the unit length of fabric by an average worker?

3.64 The price of a certain commodity in the first week of January is 400 g per rupee; it is 600 g per rupee in the second week and 500 g per rupee in the third week. Is it correct to say that the average price is 500 g per rupee? Verify.

3.65 Find the missing information in the following table:

	A	B	C	Combined
Number	10	8	—	24
Mean	20	—	6	15
Geometric Mean	10	7	—	8.397

[Delhi Univ., BCom (Hons), 1998]

3.66 During a period of decline in stock market prices, a stock is sold at Rs 50 per share on one day, Rs 40 on the next day, and Rs 25 on the third day.

(a) If an investor bought 100, 120, and 180 shares on the respective three days, find the average price paid per share.

(b) If the investor bought Rs 1000 worth of shares on each of the three days, find the average price paid per share. [Delhi Univ., BA (Hons Econ.), 1998]

Hints and Answers

3.49 (a) $\mu = \frac{\sum x_i}{N} = \frac{29.2}{3} = 9.73$ per cent

(b) $\mu_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{1,50,800}{18,000} = 8.37$ per cent

(c) The weighted mean of 8.37 per cent considering sales (in Rs) as weights is the correct mean profit margin. Percentages should never be averaged without being weighted.

3.50 $\mu = \frac{\sum x_i}{N} = \frac{196}{10} = 19.6$

$$\text{Med} = \frac{\left(\frac{n}{2}\right) + \left(\frac{n}{2} + 1\right)}{2} = \frac{5\text{th} + 6\text{th}}{2} = 20.0$$

M_0 = most frequent value = 20

(a) Median is best used as the 'typical' value because of the skewness in the distribution of values

(b) $Q_1 = x_{(n/4) + (1/2)} = x_{(10/4) + (1/2)} = x_{3.0} = 17$

$$P_{30} = x_{(3n/10) + (1/2)} = x_{3.5} = 17 + 0.5 (20 - 17) = 18.5$$

3.51 Listing the data in ascending order:

A : 14 15 16 16 17 20 30 31 53

B : 17 18 21 22 23 26 27 28 39

C : 15 16 16 17 18 19 20 31 42

$\bar{x}_A = 212/9 = 23.56$; Med (A) = 17

$\bar{x}_B = 221/9 = 24.56$; Med (B) = 23

$\bar{x}_C = 194/9 = 21.56$; Med (C) = 18

Since medians are the fifth observation in each data set, therefore design B is best because both the mean and median are highest.

3.52 Arrange the marks in statistics into following class intervals:

Marks : 0–20 20–40 40–60 60–80 80–100 100–120 120–140

No. of students : 4 26 22 10 9 6 3

Mean = 56, Median = 49.09, and Mode = 36.92

3.53 Principal amount, A = Rs 1,00,000; $r = 10$ and $n = 6$;

$$P_n = A \left(1 + \frac{r}{100}\right)^n = 1,00,000 \left(1 + \frac{10}{100}\right)^6$$

Taking log P_n and then antilog, we get $P_n = \text{Rs } 1,77,2000$

3.54 Median = Rs 679.20; Limits of income of central 50% of the employees = Rs 626.7 to Rs 747.7

3.56 Prepare a frequency distribution as follows:

Weekly Wages (Rs)	Frequency (f)	Bonus Paid (x)	Weekly Wages (Rs)	Frequency (f)	Bonus Paid (x)
61–75	3	10	106–120	5	25
76–90	4	15	121–135	7	30
91–105	5	20	136–150	6	35

Average bonus paid = $\frac{\sum fx}{n} = \frac{375}{30} = \text{Rs } 24.5$

- 3.57** Given $n_1 + n_2 = 100$, $\bar{x}_{12} = 5000$, $\bar{x}_1 = 5200$ and $\bar{x}_2 = 4200$

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$\text{or } 5000 = \frac{n_1(5200) + n_2(4200)}{n_1 + n_2}$$

$$= \frac{n_1(5200) + (100 - n_1)(4200)}{100}$$

$$1000 n_1 = 80,000 \text{ or } n_1 = 80 \text{ and } n_2 = 100 - n_1 = 20$$

- 3.58** The number of persons to be retrenched from the lower group are: $(20,000 \times 10) \div 100 = 200$. Eighty of these will be in the 15–20 age group and the rest 120 ($= 200 - 80$) in the 20–25 age group.

The persons to be absorbed in other branches $= (2,000 \times 10) \div 100 = 800$. These persons belong to the following age groups:

Age Group	No. of Persons
20–25	(250–120)
25–30	—
30–35	—
35–40	(287–242)
	800

Those who will be retiring $(2,000 \times 10) \div 100 = 200$ and these persons belong to highest age group as shown below:

Age Group	No. of Persons
65–70	—
60–65	—
55–60	—
50–55	(150–70)
	200

Age limits of persons who are retained:

Age Group	No. of Persons
35–40	242
40–45	220
45–50	268
50–55	70
	800

Average age of those retained: 43.54 years.

Output (in units)	Frequency (f)	Bonus (Rs)	Output (in units)	Frequency (f)	Bonus (Rs)
70–74	3	40	90–94	7	70
75–79	5	45	95–99	6	80
80–84	15	50	100–104	2	100
85–89	12	60			

(a) Average bonus/worker for quarter, $\bar{x} = \Sigma fx/n = 2,985/50 = \text{Rs } 59.7$

(b) Total quarterly bonus paid $= \text{Rs } 59.7 \times 50 = \text{Rs } 2,985$

(c) Average output/worker, $\bar{x} = 86.1$ units

- 3.60** Year : 1 2 3 4 5 6 7 8 9 10
Growth rate : 2 2.5 3 4 5 6 7 8 9 10
Value at the end of year x : 102 102.5 103 104 105 106 107 108 109 110
 $G.M. = \text{Antilog} (\Sigma \log x \div n) = \text{Antilog} (20.237 \div 10) = 105.6$

Average growth rate $= 105.6 - 100 = 5.6$ per cent

$$\text{3.61 H.M.} = n \sqrt[n]{\left(\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3}\right)} = \sqrt[3]{\left(\frac{1}{48} + \frac{1}{40} + \frac{1}{32}\right)} = 38.98 \text{ km per hour}$$

- 3.62** Number of employees belonging to the top 25% of the earners is $= (25 \div 100) \times 500 = 125$ and the distribution of these top earners is as follows:

Income (Rs)	Frequency
25,000 and above	10
20,000–25,000	70
15,000–20,000	45

80 persons have income in the range 15,000–20,000 = Rs 500 and therefore 45 persons will have income in the range $(500 \div 80) \times 45 = 281.25$ or 281.

The top 45 earners in the income group 15,000–20,000 will have salaries ranging from (20,000–281) to 20,000, i.e. 19,719 to 20,000. Thus the distribution of top 125 persons is as follows:

Income (Rs)	Mid-value (m)	Frequency (f)	Total Income (f × m)
25,000 and above	—	10	3,00,000 (given)
20,000–25,000	22,500	70	15,75,000
19,719–20,000	19,859.5	45	8,93,677.5
		125	27,68,677.5

Hence the total income of the top 25% of earners is Rs 2,71,177.5. Contribution to the fund = 5% of 2,71,177.5 = Rs 93,598.

- 3.63** Average time per worker before training:

$$\frac{(100 \times 3) + (250 \times 4) + (150 \times 5)}{100 + 250 + 150} = \frac{2050}{500} = 4.1 \text{ hours}$$

After training the composition of workers is as follows:

Skilled workers = 100 + 250 = 350

Semi-skilled workers = 150

Unskilled workers = Nil

Average time per worker after training is:

$$\frac{(350 \times 3) + (150 \times 4)}{350 + 150} = \frac{1050 + 600}{500} = 3.3 \text{ hours}$$

Thus after 2 years 0.8 hours less would be required.

$$\text{3.64 Harmonic Mean} = n \sqrt[n]{\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c}\right)}$$

$$= 3 \sqrt[3]{\left(\frac{1}{400} + \frac{1}{600} + \frac{1}{500}\right)} = 0.486 \text{ g per rupee}$$

3.65 Mean : Let x be the mean of B. Then

$$(20 \times 10) + (8 \times x) + (6 \times 6) = (15 \times 24)$$

$$8x = 124 \text{ or } x = 15.5$$

Hence mean of B = 15.5

Geometric mean: Let x be the geometric mean of C. Then

$$(10)^{10} \times (7)^8 \times x^6 = (8.397)^{24}$$

$$10 \log 10 + 8 \log 7 + 6 \log x = 24 \log 8.397$$

$$10 + (8 \times 0.8451) + 6 \log x = 24 (0.9241)$$

$$6 \log x = 5.4176 \text{ or } x = \text{Antilog } 0.9029 = 7.997$$

Hence, geometric mean of C is 7.997.

$$\begin{aligned}\text{3.66 Average price paid per share} &= \frac{\sum wx}{\sum w} = \frac{14,300}{400} \\ &= 35.75.\end{aligned}$$

Case Studies

Case 3.1: Kanta Bread

'Kanta Bread' company is manufacturing bread. The selling price of bread is fixed by the government. In view of increasing raw material prices the only way to sustain the profit level is by efficient control over the production process to reduce the cost of production. A large quantity of production shows small savings per unit of production and results in a significant amount cumulatively. Among various factors affecting the cost of production, the yield of the finished product obtained from a given quantity of raw materials is of considerable significance.

The process of manufacturing bread involves mixing of wheat flour (maida) with the required quantity of water and other ingredients as per the standard formula. The mixed bulk dough is subjected to yeast fermentation for a given period. Then this bulk quantity of dough is transferred to a machine called Divider, which divides the dough mechanically into small pieces of required weight, which are individually further processed and baked to get the finished product called bread.

For manufacturing a 800 g loaf of bread the required weight of dough piece coming from the Divider should be between 880 g and 885 g. This weight is termed as *dividing weight*. There is provision on the Divider machine to set the required dough piece weight. The smallest division (or increment) in weight that can be set on this machine is 5 g. As a measure of safety, 885 g is the weight that is usually set on the machine.

The performance of the machine, with respect to the accuracy of the weight of dough pieces delivered by it, is dependent on various factors such as the level of lubrication and the consistency of the dough. But these variations are less and the weight variation would be within ± 30 g of the set weight for a good Divider. Larger variations than this in weight occur only if the dough dividing mechanism (called Divider Head) has undergone considerable wear and tear.

The weight variation in the dough pieces from the Divider results in weight variation in the final product (bread). But the Bread weight, which is declared as 800 g is covered by the Weights and Measures (Packaged Commodities) Act, 1976. As per the rules of this Act, the average net weight of a random sample (Sample size is 20 bread) should not be less than the declared weight. The maximum permissible error in relation to the quantity

contained in the individual package is 6 per cent of the declared weight. Also, it is stipulated that the number of individual packages showing an error in deficiency of weight greater than the maximum permissible error (i.e., 6%) should not be more than 5 per cent of the packages drawn as samples. The sample size as per the rules as applicable to bread is 20 individual packages. If the product does not confirm to the rules, then punitive action can be taken against the manufacturer.

In order to ensure that the rules under this Act are not violated, the rules are applied to the weights of the divided dough piece itself. Thus in terms of these rules it is required:

- (i) that the average weight of a random sample of 20 divided dough pieces should be 880 g to 885 g.
- (ii) that the weight of an individual dough piece should not be less than 832 g (i.e., 885 g - 6% = 832 g).
- (iii) That not more than one piece out of 20 (i.e., 5%) should have a weight less than 832 g.

Thus the performance of the Divider should confirm to these standards. However it was complained by the production department that the weight variation at the Divider is very large and hence in order that the law is not violated, they are forced to set the Divider at a higher dividing weight than the normally required weight setting of 885 g, resulting in less on yield of the product.

Questions for Discussion

1. What is the average weight of dough pieces delivered by the Divider with the weight setting at 885 g?
2. What is the percentage of dough pieces with weight less than 832 g, if any?
3. What should be the lowest dividing weight setting on the Divider, so that the number of dough pieces with weight less than 832 g are less than 5 per cent, as stipulated by the law, (if already not so)?
4. The monetary loss to the company on account of setting the higher dividing weight as decided in Q.3 above could be calculated, which would help in deciding whether the Divider head should be replaced or not.

Data: The output of the Divider is 1000 pieces per hour. The weight of 20 dough pieces per hour. The weight of 20 dough pieces were taken using an accurate weighing scale. (Recording this weight is a routine quality control

check). Ten sets of such dough weight on the machine were taken at different times. All the individual observations on these ten sets were pooled together to get the population of data with total 200 observations, given below:

SNo.	Dividing Weight								
1.	875	41.	895	81.	836	121.	912	161.	870
2.	870	42.	910	82.	823	122.	912	162.	895
3.	852	43.	907	83.	910	123.	885	163.	870
4.	880	44.	912	84.	889	124.	895	164.	910
5.	909	45.	890	85.	897	125.	840	165.	910
6.	909	46.	895	86.	885	126.	860	166.	885
7.	893	47.	862	87.	892	127.	866	167.	920
8.	875	48.	875	88.	886	128.	875	168.	877
9.	830	49.	895	89.	886	129.	868	169.	909
10.	859	50.	867	90.	897	130.	861	170.	915
11.	827	51.	910	91.	897	131.	873	171.	920
12.	907	52.	900	92.	880	132.	893	172.	884
13.	909	53.	880	93.	870	133.	883	173.	910
14.	910	54.	900	94.	920	134.	878	174.	925
15.	915	55.	910	95.	927	135.	873	175.	900
16.	920	56.	897	96.	930	136.	893	176.	875
17.	905	57.	875	97.	925	137.	825	177.	895
18.	890	58.	905	98.	915	138.	830	178.	930
19.	925	59.	975	99.	875	139.	845	179.	863
20.	900	60.	909	100.	890	140.	830	180.	913
21.	895	61.	890	101.	890	141.	840	181.	903
22.	875	62.	880	102.	923	142.	835	182.	843
23.	903	63.	890	103.	828	143.	830	183.	878
24.	863	64.	825	104.	825	144.	860	184.	883
25.	913	65.	845	105.	910	145.	855	185.	878
26.	903	66.	875	106.	866	146.	860	186.	897
27.	893	67.	890	107.	825	147.	835	187.	916
28.	878	68.	892	108.	830	148.	886	188.	907
29.	878	69.	821	109.	845	149.	888	189.	912
30.	883	70.	826	110.	830	150.	890	190.	895
31.	897	71.	904	111.	855	151.	878	191.	885
32.	916	72.	915	112.	835	152.	858	192.	862
33.	813	73.	900	113.	855	153.	875	193.	879
34.	863	74.	900	114.	836	154.	858	194.	900
35.	900	75.	905	115.	860	155.	868	195.	872
36.	905	76.	885	116.	835	156.	888	196.	900
37.	898	77.	890	117.	865	157.	868	197.	905
38.	878	78.	889	118.	915	158.	865	198.	885
39.	878	79.	865	119.	930	159.	880	199.	895
40.	898	80.	845	120.	865	160.	905	200.	902

There never was in the world two opinions alike, no more than two hairs or two grains; the most universal quality is diversity.

—Michel de Montaigne

I feel like a fugitive from the law of averages.

—Bill Mauldin

Measures of Dispersion

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- provide the importance of the concept of variability (dispersion).
- measure the spread or dispersion, understand it, and identify its causes to provide a basis for action.

4.1 INTRODUCTION

Just as central tendency can be measured by a number in the form of an average, the amount of variation (dispersion, spread, or scatter) among the values in the data set can also be measured. The measures of central tendency describe that the major part of values in the data set appears to concentrate (cluster) around a central value called *average* with the remaining values scattered (spread or distributed) on either sides of that value. But these measures do not reveal how these values are dispersed (spread or scattered) on each side of the central value. The dispersion of values is indicated by the extent to which these values tend to spread over an interval rather than cluster closely around an average.

The statistical techniques to measure such dispersion are of two types:

- (a) Techniques that are used to measure the extent of variation or the deviation (also called degree of variation) of each value in the data set from a measure of central tendency, usually the mean or median. Such statistical techniques are called *measures of dispersion* (or *variation*).
- (b) Techniques that are used to measure the direction (away from uniformity or symmetry) of variation in the distribution of values in the data set. Such statistical techniques are called *measures of skewness*.

To measure the dispersion, understand it, and identify its causes is very important in statistical inference (estimation of parameter, hypothesis testing, forecasting, and so on). A small dispersion among values in the data set indicates that data are clustered closely around the mean. The mean is therefore considered representative of the data,

i.e. mean is a reliable average. Conversely, a large dispersion among values in the data set indicates that the mean is not reliable, i.e. it is not representative of the data.

Figure 4.1
Symmetrical Distributions with
Unequal Mean and Equal Standard
Deviation

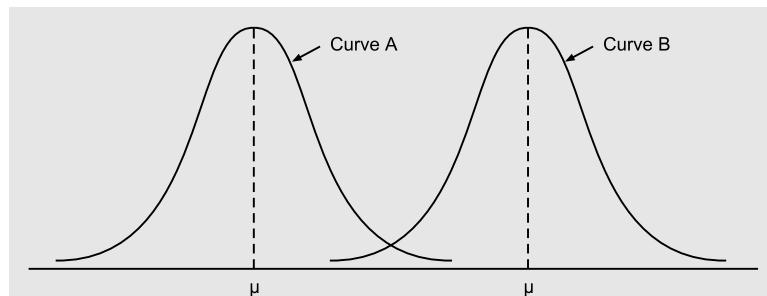
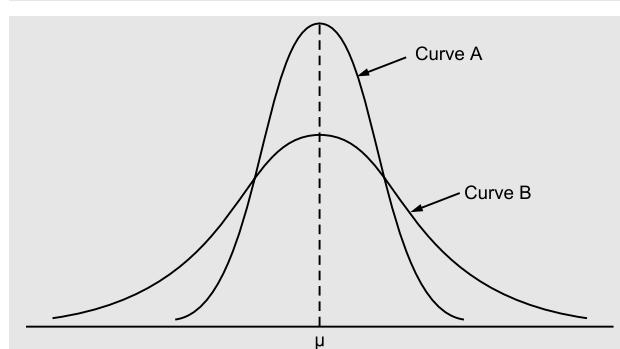


Figure 4.2
Symmetrical Distributions with
Equal Mean and Unequal Standard
Deviation



The symmetrical distribution of values in two or more sets of data may have same variation but differ greatly in terms of A.M. On the other hand, two or more sets of data may have the same A.M. values but differ in variation as shown in Fig. 4.2.

Illustration Suppose over the six-year period the net profits (in percentage) of two firms are as follows:

Firm 1 :	5.2,	4.5,	3.9,	4.7,	5.1,	5.4
Firm 2 :	7.8,	7.1,	5.3,	14.3,	11.0,	16.1

Since average amount of profit is 4.8 per cent for both firms, therefore operating results of both the firms are equally good and that a choice between them for investment purposes must depend on other considerations. However, the difference among the values is greater in Firm 2, that is, profit is varying from 5.3 to 16.1 per cent, while the net profit values of Firm 1 were varying from 3.9 to 5.4 per cent. This shows that the values in data set 2 are spread more than those in data set 1. This implies that Firm 1 has a consistent performance while Firm 2 has a highly inconsistent performance. Thus for investment purposes, a comparison of the average (mean) profit values alone should not be sufficient.

4.2 SIGNIFICANCE OF MEASURING DISPERSION

Following are some of the purposes for which measures of variation are needed.

1. **Test the reliability of an average:** Measures of variation are used to test to what extent an average represents the characteristic of a data set. If the variation is small, that is, extent of dispersion or scatter is less on each side of an average, then it indicates high uniformity of values in the distribution and the average represents an individual value in the data set. On the other hand, if the variation is large, then it indicates a lower degree of uniformity in values in the data set, and the average may be unreliable. No variation indicates perfect uniformity and, therefore, values in the data set are identical.
2. **Control the variability:** Measuring variation helps to identify the nature and causes of variation. Such information is useful in controlling the variations. According to Spurr and Bonini, 'In matters of health, variations in, body temperature, pulse beat and

blood pressure are the basic guides to diagnosis. Prescribed treatment is designed to control their variation. In industrial production, efficient operation requires control of quality variation, the causes of which are sought through inspection and quality control programmes.' In social science, the measurement of 'inequality' of distribution of income and wealth requires the measurement of variability.

3. **Compare two or more sets of data with respect to their variability:** Measures of variation help in the comparison of the spread in two or more sets of data with respect to their uniformity or consistency. For example, (i) the measurement of variation in share prices and their comparison with respect to different companies over a period of time requires the measurement of variation, (ii) the measurement of variation in the length of stay of patients in a hospital every month may be used to set staffing levels, number of beds, number of doctors, and other trained staff, patient admission rates, and so on.
4. **Facilitate the use of other statistical techniques:** Measures of variation facilitate the use of other statistical techniques such as correlation and regression analysis, hypothesis testing, forecasting, quality control, and so on.

4.2.1 Essential Requisites for a Measure of Variation

The essential requisites for a good measure of variation are listed below. These requisites help in identifying the merits and demerits of individual measure of variation.

- (i) It should be rigidly defined.
- (ii) It should be based on all the values (elements) in the data set.
- (iii) It should be calculated easily, quickly, and accurately.
- (iv) It should not be unduly affected by the fluctuations of sampling and also by extreme observations.
- (v) It should be amenable to further mathematical or algebraic manipulations.

4.3 CLASSIFICATION OF MEASURES OF DISPERSION

The various measures of dispersion (variation) can be classified into two categories:

- (i) Absolute measures, and
- (ii) Relative measures

Absolute measures are described by a number or value to represent the amount of variation or differences among values in a data set. Such a number or value is expressed in the same unit of measurement as the set of values in the data such as rupees, inches, feet, kilograms, or tonnes. Such measures help in comparing two or more sets of data in terms of absolute magnitude of variation, provided the variable values are expressed in the same unit of measurement and have almost the same average value.

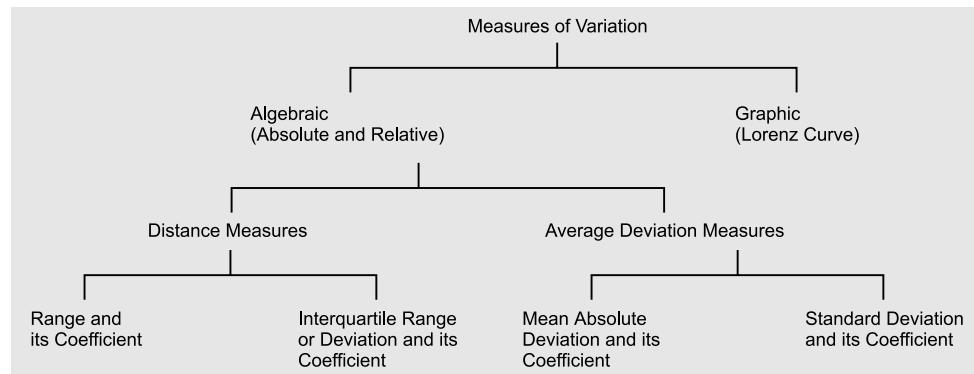
The *relative measures* are described as the ratio of a measure of absolute variation to an average and is termed as *coefficient of variation*. The word 'coefficient' means a number that is independent of any unit of measurement. While computing the relative variation, the average value used as base should be the same from which the absolute deviations were calculated.

Another classification of the measures of variation is based on the method employed for their calculations:

- (i) Distance measures, and
- (ii) Average deviation measures

The *distance measures* describe the spread or dispersion of values of a variable in terms of difference among values in the data set. The *average deviation measures* describe the average deviation for a given measure of central tendency.

The above-mentioned classification of various measures of dispersion (variation) may be summarized as shown below:



4.4 DISTANCE MEASURES

As mentioned above, two distance measures discussed in this section are namely:

- (i) Range, and
- (ii) Interquartile deviation

4.4.1 Range

Range: A measure of variability, defined to be the difference between the largest and lowest values in the data set.

The range is the most simple measure of dispersion and is based on the location of the largest and the smallest values in the data. Thus, the *range is defined to be the difference between the largest and lowest observed values in a data set*. In other words, it is the length of an interval which covers the highest and lowest observed values in a data set and thus measures the dispersion or spread within the interval in the most direct possible way.

$$\begin{aligned} \text{Range (R)} &= \text{Highest value of an observation} - \text{Lowest value of an observation} \\ &= H - L \end{aligned} \quad (4-1)$$

For example, if the smallest value of an observation in the data set is 160 and largest value is 250, then the range is $250 - 160 = 90$.

For grouped frequency distributions of values in the data set, the range is the difference between the upper class limit of the last class and the lower class limit of first class. In this case, the range obtained may be higher than as compared to ungrouped data because of the fact that the class limits are extended slightly beyond the extreme values in the data set.

Coefficient of Range

The relative measure of range, called the coefficient of range is obtained by applying the following formula:

$$\text{Coefficient of range} = \frac{H - L}{H + L} \quad (4-2)$$

Example 4.1: The following are the sales figures of a firm for the last 12 months

Months :	1	2	3	4	5	6	7	8	9	10	11	12
Sales												
(Rs '000) :	80	82	82	84	84	86	86	88	88	90	90	92

Calculate the range and coefficient of range for sales.

Solution: Given that $H = 92$ and $L = 80$. Therefore

$$\text{Range} = H - L = 92 - 80 = 12$$

$$\text{and} \quad \text{Coefficient of range} = \frac{H - L}{H + L} = \frac{92 - 80}{92 + 80} = \frac{12}{172} = 0.069$$

Example 4.2: The following data show the waiting time (to the nearest 100th of a minute) of telephone calls to be matured:

<i>Waiting Time</i>	<i>Frequency (Minutes)</i>	<i>Waiting Time</i>	<i>Frequency (Minutes)</i>
0.10–0.35	6	0.88–1.13	8
0.36–0.61	10	1.14–1.39	4
0.62–0.87	8		

Calculate the range and coefficient of range.

Solution: Given that, $H = 1.39$ and $L = 0.10$. Therefore

$$\text{Range} = H - L = 1.39 - 0.10 = 1.29 \text{ minutes}$$

$$\text{and } \text{Coefficient of Range} = \frac{H - L}{H + L} = \frac{1.39 - 0.10}{1.39 + 0.10} = \frac{1.29}{1.49} = 0.865$$

Advantages, Disadvantages and Applications of Range The major advantages and disadvantages of range may be summarized as follows:

Advantages

- (i) It is independent of the measure of central tendency and easy to calculate and understand.
- (ii) It is quite useful in cases where the purpose is only to find out the extent of extreme variation, such as industrial quality control, temperature, rainfall, and so on.

Disadvantages

- (i) The calculation of range is based on only two values—largest and smallest in the data set and fail to take account of any other observations.
- (ii) It is largely influenced by two extreme values and completely independent of the other values. For example, range of two data sets $\{1, 2, 3, 7, 12\}$ and $\{1, 1, 1, 12, 12\}$ is 11, but the two data sets differ in terms of overall dispersion of values
- (iii) Its value is sensitive to changes in sampling, that is, different samples of the same size from the same population may have widely different ranges.
- (iv) It cannot be computed in case of open-ended frequency distributions because no highest or lowest value exists in open-ended class.
- (v) It does not describe the variation among values in the data between two extremes.

For example, each of the following set of data

Set 1 :	9	21	21	21	21	21	21	21
Set 2 :	9	9	9	9	21	21	21	21
Set 3 :	9	10	12	14	15	19	20	21

has a range of $21 - 9 = 12$, but the variation of values is quite different in each case between the highest and lowest values.

Applications of Range

- (i) *Fluctuation in share prices:* The range is useful in the study of small variations among values in a data set, such as variation in share prices and other commodities that are very sensitive to price changes from one period to another.
- (ii) *Quality control:* It is widely used in industrial quality control. Quality control is exercised by preparing suitable *control charts*. These charts are based on setting an upper control limit (range) and a lower control limit (range) within which produced items shall be accepted. The variation in the quality beyond these ranges requires necessary correction in the production process or system.
- (iii) *Weather forecasts:* The concept of range is used to determine the difference between maximum and minimum temperature or rainfall by meteorological departments to announce for the knowledge of the general public.

4.4.2 Interquartile Range or Deviation

Interquartile range: A measure of variability, defined to be the difference between the quartiles Q_3 and Q_1 .

The limitations or disadvantages of the range can partially be overcome by using another measure of variation which measures the spread over the middle half of the values in the data set so as to minimise the influence of outliers (extreme values) in the calculation of range. Since a large number of values in the data set lie in the central part of the frequency distribution, therefore it is necessary to study the **Interquartile Range** (also called midspread). To compute this value, the entire data set is divided into four parts each of which contains 25 per cent of the observed values. The quartiles are the highest values in each of these four parts. The *interquartile range* is a measure of dispersion or spread of values in the data set between the third quartile, Q_3 and the first quartile, Q_1 . In other words, the *interquartile range or deviation* (IQR) is the range for the middle 50 per cent of the data. The concept of IQR is shown in Fig. 4.3:

$$\text{Interquartile range (IQR)} = Q_3 - Q_1 \quad (4-3)$$

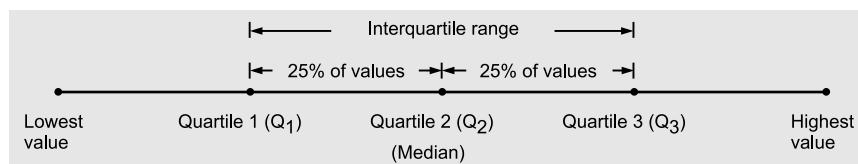
Half the distance between Q_1 and Q_3 is called the *semi-interquartile range* or the *quartile deviation* (QD).

$$\text{Quartile deviation (QD)} = \frac{Q_3 - Q_1}{2} \quad (4-4)$$

The median is not necessarily midway between Q_1 and Q_3 , although this will be so for a symmetrical distribution. The median and quartiles divide the data into equal numbers of values but do not necessarily divide the data into equally wide intervals.

As shown above the quartile deviation measures the average range of 25 per cent of the values in the data set. It represents the spread of all observed values because its value is computed by taking an average of the middle 50 per cent of the observed values rather than of the 25 per cent part of the values in the data set.

Figure 4.3
Interquartile Range



In a non-symmetrical distribution, the two quartiles Q_1 and Q_3 are at equal distance from the median, that is, $\text{Median} - Q_1 = Q_3 - \text{Median}$. Thus, $\text{Median} \pm \text{Quartile Deviation}$ covers exactly 50 per cent of the observed values in the data set.

A smaller value of quartile deviation indicates high uniformity or less variation among the middle 50 per cent observed values around the median value. On the other hand, a high value of quartile deviation indicates large variation among the middle 50 per cent observed values.

Coefficient of Quartile Deviation

Since quartile deviation is an absolute measure of variation, therefore its value gets affected by the size and number of observed values in the data set. Thus, the Q.D. of two or more than two sets of data may differ. Due to this reason, to compare the degree of variation in different sets of data, we compute the relative measure corresponding to Q.D., called the *coefficient of Q.D.*, and it is calculated as follows:

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \quad (4-5)$$

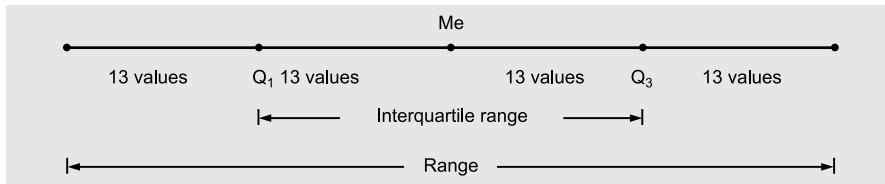
Example 4.3: Following are the responses from 55 students to the question about how much money they spent every day.

55	60	80	80	80	85	85	85	90	90	90
90	92	94	95	95	95	95	100	100	100	100
100	100	105	105	105	105	109	110	110	110	110
112	115	115	115	115	115	120	120	120	120	120
124	125	125	125	130	130	140	140	140	145	150

Calculate the range and interquartile range and interpret your result.

Solution: The median of the given values in the data set is the $(55+1)/2 = 28$ th value which is 105. From this middle value of 105, there are 27 values at or below 105 and another 27 at or above 105.

The lower quartile $Q_1 = (27+1)/2 = 14$ th value from bottom of the data i.e. $Q_1 = 94$ and upper quartile is the 14th value from the top, i.e. $Q_3 = 120$. The 55 values have been partitioned as follows:



The interquartile range, $IQR = 120 - 94 = 26$ while the range is $R = 150 - 55 = 95$. The middle 50% of the data fall in relatively narrow range of only Rs 26. This means responses are more densely clustered near the centre of the data and more spread out towards the extremes. For instance, lowest 25% of the students had responses ranging over 55 to 94, i.e. Rs 39, while the next 25% had responses ranging over 94 to 105, i.e. only Rs 11. Similarly, the third quarter had responses from 105 to 110, i.e. only Rs 5, while the top 25% had responses in the interval (120 to 150), i.e. Rs 30.

The median and quartiles divide the data into equal number of values but not necessarily divide the data into equally wide intervals.

Example 4.4: Use an appropriate measure to evaluate the variation in the following data:

Farm Size (acre)	No. of Farms	Farm Size (acre)	No. of Farms
below 40	394	161–200	169
41–80	461	201–240	113
81–120	391	241 and above	148
121–160	334		

Solution: Since the frequency distribution has open-end class intervals on the two extreme sides, therefore Q.D. would be an appropriate measure of variation. The computation of Q.D. is shown in Table 4.1.

Table 4.1 Calculations of Quartile Deviation

Farm Size (acre)	No. of Farms	Cumulative Frequency (cf) (less than)
below 40	394	394
41–80	461	855 ← Q_1 class
81–120	391	1246
121–160	334	1580 ← Q_3 class
161–200	169	1749
201–240	113	1862
241 and above	148	2010
		2010

$$Q_1 = \text{Value of } (n/4)\text{th observation} = 2010 \div 4 \text{ or } 502.5\text{th observation}$$

This observation lies in the class 41–80. Therefore

$$\begin{aligned} Q_1 &= l + \frac{(n/4) - cf}{f} \times h \\ &= 41 + \frac{502.5 - 394}{461} \times 40 = 41 + 9.41 = 50.41 \text{ acres} \end{aligned}$$

$$Q_3 = \text{Value of } (3n/4)\text{th observation} = (3 \times 2010) \div 4 \text{ or } 1507.5\text{th observation}$$

This observation lies in the class 121–160. Therefore

$$\begin{aligned} Q_3 &= l + \frac{(3n/4) - cf}{f} \times h \\ &= 121 + \frac{1507.5 - 1246}{334} \times 40 = 121 + 31.31 = 152.31 \text{ acres} \end{aligned}$$

Thus, the quartile deviation is given by

$$Q.D. = \frac{Q_3 - Q_1}{2} = \frac{152.31 - 50.41}{2} = 50.95 \text{ acres}$$

$$\text{and} \quad \text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{50.95}{202.72} = 0.251$$

Advantages and Disadvantages of Quartile Deviation The major advantages and disadvantages of quartile deviation are summarized as follows:

Advantages

- (i) It is not difficult to calculate but can only be used to evaluate variation among observed values within the middle of the data set. Its value is not affected by the extreme (highest and lowest) values in the data set.
- (ii) It is an appropriate measure of variation for a data set summarized in open-ended class intervals.
- (iii) Since it is a positional measure of variation, therefore it is useful in case of erratic or highly skewed distributions, where other measures of variation get affected by extreme values in the data set.

Disadvantages

- (i) The value of Q.D. is based on the middle 50 per cent observed values in the data set, therefore it cannot be considered as a good measure of variation as it is not based on all the observations.
- (ii) The value of Q.D. is very much affected by sampling fluctuations.
- (iii) The Q.D. has no relationship with any particular value or an average in the data set for measuring the variation. Its value is not affected by the distribution of the individual values within the interval of the middle 50 per cent observed values.

Conceptual Questions 4A

- Explain the term variation. What does a measure of variation serve? In the light of these, comment on some of the well-known measures of variation.
[Delhi Univ., MBA, 2001]
- What are the requisites of a good measure of variation?
- Explain how measures of central tendency and measures of variation are complementary to each other in the context of analysis of data.
- Distinguish between absolute and relative measures of variation. Give a broad classification of the measures of variation.

- (a) Critically examine the different methods of measuring variation.
(b) Explain with suitable examples the term ‘variation’. Mention some common measures of variation and describe the one which you think is the most important.
[Delhi Univ., MBA, 1998]
- Explain and illustrate how the measures of variation afford a supplement to the information about frequency distribution furnished by averages.
[Delhi Univ., MBA, 1999]
- What do you understand by ‘coefficient of variation’? Discuss its importance in business problems.

Self-Practice Problems 4A

- The following are the prices of shares of a company from Monday to Saturday:

Days	Price (Rs)	Days	Price (Rs)
Monday	200	Thursday	160
Tuesday	210	Friday	220
Wednesday	208	Saturday	250

Calculate the range and its coefficient.

- 4.2** The day's sales figures (in Rs) for the last 15 days at Nirula's ice-cream counter, arranged in ascending order of magnitude, are recorded as follows: 2000, 2000, 2500, 2500, 2500, 3500, 4000, 5300, 9000, 12,500, 13,500, 24,500, 27,100, 30,900, and 41,000. Determine the range and middle 50 per cent range for this sample data.
- 4.3** The following distribution shows the sales of the fifty largest companies for a recent year:

Sales (Million of rupees)	Number of Companies
0–9	18
10–19	19
20–29	6
30–39	2
40–49	5

Calculate the coefficient of range

- 4.4** You are given the frequency distribution of 292 workers of a factory according to their average weekly income.

Weekly Income (Rs)	No. of Workers	Weekly Income (Rs)	No. of Workers
Below 1350	8	1450–1470	22
1350–1370	16	1470–1490	15
1370–1390	39	1490–1510	15
1390–1410	58	1510–1530	9
1410–1430	60	1530 and above	10
1430–1450	40		

Calculate the quartile deviation and its coefficient from the above mentioned data.

[Kurukshetra Univ., MBA, 1998]

- 4.5** You are given the data pertaining to kilowatt hours of electricity consumed by 100 persons in a city.

Consumption (kilowatt hour)	No. of Users
0–10	6
10–20	25
20–30	36
30–40	20
40–50	13

Calculate the range within which the middle 50 per cent of the consumers fall.

- 4.6** The following sample shows the weekly number of road accidents in a city during a two-year period:

Number of Accidents	Frequency	Number of Accidents	Frequency
0–4	5	25–29	9
5–9	12	30–34	4
10–14	32	35–39	3
15–19	27	40–44	1
20–24	11		

Find the interquartile range and coefficient of quartile deviation.

- 4.7** A City Development Authority subdivided the available land for housing into the following building lot sizes:

Lot Size (Square meters)	Frequency
Below 69.44	19
69.44–104.15	25
104.16–208.32	42
208.33–312.49	12
312.50–416.65	5
416.66 and above	17

Find the interquartile range and quartile deviation.

- 4.8** The cholera cases reported in different hospitals of a city in a rainy season are given below:

Calculate the quartile deviation for the given distribution and comment upon the meaning of your result.

Age Group (Years)	Frequency	Age Group (Years)	Frequency
Less than 1	15	25–35	132
1–5	113	35–45	65
5–10	122	45–65	46
10–15	91	65 and above	15
15–25	229		

Hints and Answers

- 4.1** Range = Rs 90, Coefficient of range = 0.219

- 4.2** Range = Rs 39,000;

Middle 50%, R = $P_{75} - P_{25}$

$$\begin{aligned}
 &= x_{(75n/100) + (1/2)} - x_{(25n/100) + (1/2)} \\
 &= x_{(11.25 + 0.50)} - x_{(3.75 + 0.50)} \\
 &= x_{11.75} - x_{4.25} \\
 &= (13,500 + 8250) - (2500 + 00) \\
 &= 19,250
 \end{aligned}$$

Here $x_{11.75}$ is the interpolated value for the 75% of the

distance between 11th and 12th ordered sales amount. Similarly, $x_{4.25}$ is the interpolated value for the 25% of the distance between 4th and 5th order sales amount.

- 4.3** Coefficient of range = 1

- 4.4** Quartile deviation = 27.76; Coeff. of Q.D. = 0.020; $Q_1 = 1393.48$; $Q_3 = 1449$

- 4.5** $Q_3 - Q_1 = 34 - 17.6 = 16.4$

- 4.6** $Q_3 - Q_1 = 30.06$; Coefficient of Q.D. = 0.561

- 4.7** $Q_3 - Q_1 = 540.26$; Q.D. = 270.13

- 4.8** Q.D. = 10 years

4.5 AVERAGE DEVIATION MEASURES

The range and quartile deviation indicate overall variation in a data set, but do not indicate spread or scatteredness around the central value (i.e. mean, median or mode). However, to understand the nature of distribution of values in the data set, we need to measure the 'spread' of values around the mean to indicate how representative the mean is.

In this section, we shall discuss two more measures of dispersion to measure the mean (or average) amount by which all values in a data set (population or sample) vary from their mean. These measures deal with the average deviation from some measure of central tendency—usually mean or median. These measures are:

- (a) Mean Absolute Deviation or Average Deviation
- (b) Variance and Standard Deviation

4.5.1 Mean Absolute Deviation

Since two measures of variation, range and quartile deviation, discussed earlier do not show how values in a data set are scattered about a central value or disperse themselves throughout the range, therefore it is quite reasonable to measure the variation as a degree (amount) to which values within a data set deviate from either mean or median.

The mean of deviations of individual values in the data set from their actual mean is always zero so such a measure (zero) would be useless as an indicator of variation. This problem can be solved in two ways:

- (i) Ignore the signs of the deviations by taking their absolute value, or
- (ii) Square the deviations because the square of a negative number is positive.

Since the absolute difference between a value x_i of an observation from A.M. is always a positive number, whether it is less than or more than the A.M., therefore we take the absolute value of each such deviation from the A.M. (or median). Taking the average of these deviations from the A.M., we get a measure of variation called the *mean absolute deviation* (MAD). In general, the mean absolute deviation is given by

$$\text{MAD} = \frac{1}{N} \sum_{i=1}^N |x_i - \mu|, \quad \text{for a population} \quad (4-6)$$

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|, \quad \text{for a sample}$$

where $||$ indicates the absolute value. That is, the signs of deviations from the mean are disregarded.

For a grouped frequency distribution, MAD is given by

$$\text{MAD} = \frac{\sum_{i=1}^n f_i |x_i - \bar{x}|}{\sum f_i} \quad (4-7)$$

Formulae (4-6) and (4-7), in different contexts, indicate that the MAD provides a useful method of comparing the relative tendency of values in the distribution to scatter around a central value or to disperse themselves throughout the range.

While calculating the mean absolute deviation, the median is also considered for computing because the sum of the absolute values of the deviations from the median is smaller than that from any other value. However, in general, arithmetic mean is used for this purpose.

If a frequency distribution is symmetrical, then A.M. and median values coincide and the same MAD value is obtained. In such a case $\bar{x} \pm \text{MAD}$ provides a range in which 57.5 per cent of the observations are included. Even if the frequency distribution

is moderately skewed, the interval $\bar{x} \pm \text{MAD}$ includes the same percentage of observations. This shows that more than half of the observations are scattered within one unit of the MAD around the arithmetic mean.

The MAD is useful in situations where occasional large and erratic deviations are likely to occur. The standard deviation, which uses the squares of these large deviations, tends to over-emphasize them.

Coefficient of MAD

The relative measure of mean absolute deviation (MAD) called the *coefficient of MAD* is obtained by dividing the MAD by a measure of central tendency (arithmetic mean or median) used for calculating the MAD. Thus

$$\text{Coefficient of MAD} = \frac{\text{Mean absolute deviation}}{\bar{x} \text{ or Me}} \quad (4-8)$$

If the value of relative measure is desired in percentage, then

$$\text{Coefficient of MAD} = \frac{\text{MAD}}{\bar{x} \text{ or Me}} \times 100$$

Example 4.5: The number of patients seen in the emergency ward of a hospital for a sample of 5 days in the last month were: 153, 147, 151, 156 and 153. Determine the mean deviation and interpret.

Solution: The mean number of patients is, $\bar{x} = (153 + 147 + 151 + 156 + 153)/5 = 152$. Below are the details of the calculations of MAD using formula (4-6).

Numer of Patients (x)	Absolute Deviation	
	$x - \bar{x}$	$ x - \bar{x} $
153	$153 - 152 = 1$	1
147	$147 - 152 = -5$	5
151	$151 - 152 = -1$	1
156	$156 - 152 = 4$	4
153	$153 - 152 = 1$	1
		12

$$\text{MAD} = \frac{1}{n} \sum |x - \bar{x}| = \frac{12}{5} = 2.4 \approx 3 \text{ patients (approx)}$$

The mean absolute deviation is 3 patients per day. The number of patients deviate on the average by 3 patients from the mean of 152 patients per day.

Example 4.6: Calculate the mean absolute deviation and its coefficient from median for the following data

Year	Sales (Rs thousand)	
	Product A	Product B
1996	23	36
1997	41	39
1998	29	36
1999	53	31
2000	38	47

Solution: The median sales (Me) of the two products A and B is $Me = 38$ and $Me = 36$, respectively. The calculations of MAD in both the cases are shown in Table 4.2.

Table 4.2 Calculations of MAD

Product A		Product B	
Sales (x)	$ x - Me = x - 38 $	Sales (x)	$ x - Me = x - 36 $
23	15	31	05
29	09	36	00
38	00	36	00
41	03	39	03
53	15	47	11
$n = 5$	$\sum x - Me = 42$	$n = 5$	$\sum x - Me = 19$

$$\text{Product A: } \text{MAD} = \frac{1}{n} \sum |x - Me| = \frac{42}{5} = 8.4$$

$$\text{Coefficient of MAD} = \frac{\text{MAD}}{\text{Me}} = \frac{8.4}{38} = 0.221$$

$$\text{Product B: } \text{MAD} = \frac{1}{n} \sum |x - Me| = \frac{19}{5} = 3.8$$

$$\text{Coefficient of MAD} = \frac{\text{MAD}}{\text{Me}} = \frac{3.8}{36} = 0.106$$

Example 4.7: Find the mean absolute deviation from mean for the following frequency distribution of sales (Rs in thousand) in a co-operative store.

Sales	:	50–100	100–150	150–200	200–250	250–300	300–350
Number of days :		11	23	44	19	8	7

Solution: The mean absolute deviation can be calculated by using the formula (4-6) for mean. The calculations for MAD are shown in Table 4.3. Let the assumed mean be, $A = 175$.

Table 4.3 Calculations for MAD

Sales (Rs)	Mid-Value (m)	Frequency (f)	$(m - 175)/50$ (= d)	fd	$ x - \bar{x} $ = $ x - \bar{x} $	$f x - \bar{x} $
50–100	75	11	-2	-22	104.91	1154.01
100–150	125	23	-1	-23	54.91	1262.93
150–200	175 ← A	44	0	0	4.91	216.04
200–250	225	19	1	19	45.09	856.71
250–300	275	8	2	16	95.09	760.72
300–350	325	7	3	21	145.09	1015.63
		112		11		5266.04

$$\bar{x} = A + \frac{\sum fd}{\sum f} \times h = 175 + \frac{11}{112} \times 50 = \text{Rs } 179.91 \text{ per day}$$

$$\text{MAD} = \frac{\sum f|x - \bar{x}|}{\sum f} = \frac{5266.04}{112} = \text{Rs } 47.01$$

Thus, the average sales is Rs 179.91 thousand per day and the mean absolute deviation of sales is Rs 47.01 thousand per day.

Example 4.8: A welfare organization introduced an education scholarship scheme for school going children of a backward village. The rates of scholarship were fixed as given below:

Age Group (Years)	Amount of Scholarship per Month (Rs)
5–7	300
8–10	400
11–13	500
14–16	600
17–19	700

The ages of 30 school 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, 3, 5 years respectively. Calculate mean and standard deviation of monthly scholarship. Find out the total monthly scholarship amount being paid to the students. [IGNOU, MBA, 2002]

Solution: The number of students in the age group from 5–7 to 17–19 are calculated as shown in table 4.4:

Table 4.4

Age Group (Years)	Tally Bars	Number of Students
5–7		10
8–10		8
11–13		7
14–16		3
17–19		2
		30

The calculations for mean and standard deviation are shown in Table 4.5.

Table 4.5 Calculations for Mean and Standard Deviation

Age Group (Years)	Number of Students (<i>f</i>)	Mid-value (<i>m</i>)	$d = \frac{m - A}{h} = \frac{m - 12}{3}$	<i>fd</i>	<i>fd</i> ²
5–7	10	6	-2	-20	40
8–10	8	9	-1	-8	8
11–13	7	A → 12	0	0	0
14–16	3	15	1	3	3
17–19	2	18	2	4	8
	30			-21	59

$$\text{Mean, } \bar{x} = A + \frac{\sum fd}{\sum f} \times h = 12 - \frac{21}{30} \times 3 = 12 - 2.1 = 9.9$$

$$\begin{aligned} \text{Standard deviation, } \sigma &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times h = \sqrt{\frac{59}{30} - \left(\frac{-21}{30}\right)^2} \times 3 \\ &= \sqrt{1.967 - 0.49} \times 3 = 1.2153 \times 3 = 3.6459 \end{aligned}$$

Calculations for monthly scholarship paid to 30 students are shown in Table 4.6.

Table 4.6 Calculations for Monthly Scholarship

Number of Students	Amount of Scholarship per Month (Rs)	Total Monthly Scholarship (Rs)
10	300	3000
8	400	3200
7	500	3500
3	600	1800
2	700	1400
		12,900

Advantages and Disadvantages of MAD The advantages and disadvantages of MAD are summarized below:

Advantages

- (i) The calculation of MAD is based on all observations in the distribution and shows the dispersion of values around the measure of central tendency.
- (ii) The value of MAD is easy to compute and therefore makes it popular among those users who are not even familiar with statistical methods.
- (iii) While calculating MAD, equal weightage is given to each observed value and thus it indicates how far each observation lies from either the mean or median.
- (iv) Average deviation from mean is always zero in any data set. The MAD avoids this problem by using absolute values to eliminate the negative signs.

Disadvantages

- (i) The algebraic signs are ignored while calculating MAD. If the signs are not ignored, then the sum of the deviations taken from arithmetic mean will be zero and close to zero when deviations are taken from median.
 - (ii) The value of MAD is considered to be best when deviations are taken from median. However, median does not provide a satisfactory result in case of a high degree of variability in a data set.
- Moreover, the sum of the deviations from mean (ignoring signs) is greater than the sum of the deviations from median (ignoring signs). In such a situation, computations of MAD by taking deviations from mean is also not desirable.
- (iii) The MAD is generally unwieldy in mathematical discussions.

Inspite of all these demerits, the knowledge of MAD would help the reader to understand another important measure of dispersion called the *standard deviation*.

4.5.2 Variance and Standard Deviation

Another way to disregard the signs of negative deviations from mean is to square them. Instead of computing the absolute value of each deviation from mean, we square the deviations from mean. Then the sum of all such squared deviations is divided by the number of observations in the data set. This value is a measure called **population variance** and is denoted by σ^2 (a lower-case Greek letter sigma). It is usually referred to as 'sigma squared'. Symbolically, it is written as:

$$\text{Population variance, } \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad (4-9)$$

$$= \frac{1}{N} \sum_{i=1}^N x_i^2 - (\mu)^2$$

(Deviation is taken from actual population A.M.)

$$= \frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N} \right)^2 \quad (\text{Deviation is taken from assumed A.M.})$$

where $d = x - A$ and A is any constant (also called assumed A.M.).

Since σ^2 is the average or mean of squared deviations from arithmetic mean, it is also called the *mean square average*.

The population variance is basically used to measure variation among the values of observations in a population. Thus for a population of N observations (elements) and with μ denoting the population mean, the formula for population variance is shown in Eqn. (4-9). However, in almost all applications of statistics, the data being analyzed is a sample data. As a result, population variance is rarely determined. Instead, we compute a sample variance to estimate population variance, σ^2 .

Variance: A measure of variability based on the squared deviations of the observed values in the data set about the mean value.

It was shown that if the *sum of the squared deviations* about a sample mean \bar{x} in Eqn. (4-9) is divided by n (sample size), then it invariably tends to cause the resulting estimate of σ^2 to be lower than its actual value. This undesirable condition is called *bias*. However, this *bias* in the estimation of population variance from a sample can be removed by dividing the sum of the squared deviations between the sample mean and each element in the population by $n - 1$ rather than by n . Thus, the *unbiased* sample variance denoted by s^2 is defined as follows:

$$\text{Sample variance, } s^2 = \frac{\sum(x - \bar{x})^2}{n-1} = \frac{\sum x^2}{n-1} - \frac{n \bar{x}^2}{n-1} = \frac{\Sigma x^2}{n-1} - \frac{(\Sigma x)^2}{n(n-1)} \quad (4-10)$$

The numerator $\sum(x - \bar{x})^2$ in Eqn. (4-10) is called the *total sum of squares*. This quantity measures the total variation among values in a data set (whereas the variance measures only the *average variation*). The larger the value of $\sum(x - \bar{x})^2$, the greater the variation among the values in a data set.

Standard Deviation

The numerical value of population or a sample variance is difficult to interpret because it is expressed in square units. To reach a interpretable measure of variance expressed in the units of original data, we take a positive square root of the variance, which is known as the **standard deviation** or *root-mean square deviation*. The standard deviation of population and sample is denoted by σ and s , respectively. We can think of the standard deviation as roughly the *average distance values fall from the mean*.

(a) Ungrouped Data

$$\begin{aligned} \text{Population standard deviation, } \sigma &= \sqrt{\sigma^2} = \sqrt{\frac{1}{N} \sum (x - \mu)^2} = \sqrt{\frac{1}{N} \sum x^2 - (\mu)^2} \\ &= \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N} \right)^2} \end{aligned}$$

$$\text{Sample standard deviation, } s = \sqrt{\frac{\sum x^2}{n-1} - \frac{n \bar{x}^2}{n-1}} ; \text{ where } n = \text{sample size}$$

(b) Grouped Data

$$\text{Population standard deviation, } \sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2} \times h$$

where f = frequency of each class interval

$N = \sum f$ = total number of observations (or elements) in the population

h = width of class interval

m = mid-value of each class interval

$$d = \frac{m - A}{h}, \text{ where } A \text{ is any constant (also called assumed A.M.)}$$

$$\text{Sample standard deviation, } s = \sqrt{s^2} = \sqrt{\frac{\sum f(x - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum fx^2}{n-1} - \frac{(\sum fx)^2}{n(n-1)}} \quad (4-11)$$

Remarks: 1. For any data set, MAD is always less than the σ because MAD is less sensitive to the extreme observations. Thus, when a data contains few very large observations, the MAD provides a more realistic measure of variation than σ . However σ is often used in statistical applications because it is amenable to mathematical development.

2. When sample size (n) becomes very large, $(n - 1)$ becomes indistinguishable and irrelevant.

Advantages and Disadvantages of Standard Deviation The advantages and disadvantages of the standard deviation are summarized below:

Standard deviation: A measure of variability computed by taking the positive square root of the variance.

Advantages

- (i) The value of standard deviation is based on every observation in a set of data. It is the only measure of variation capable of algebraic treatment and less affected by fluctuations of sampling as compared to other measures of variation.
- (ii) It is possible to calculate the combined standard deviation of two or more sets of data.
- (iii) Standard deviation has a definite relationship with the area under the symmetric curve of a frequency distribution. Due to this reason, standard deviation is called a *standard* measure of variation.
- (iv) Standard deviation is useful in further statistical investigations. For example, standard deviation plays a vital role in comparing skewness, correlation, and so on, and also widely used in sampling theory.

Disadvantages

- (i) As compared to other measures of variation, calculations of standard deviation are difficult.
- (ii) While calculating standard deviation, more weight is given to extreme values and less to those near mean. Since for calculating S.D., the deviations from the mean are squared, therefore large deviations when squared are proportionately more than small deviations. For example, the deviations 2 and 10 are in the ratio of 1 : 5 but their squares 4 and 100 are in the ratio of 1 : 25.

Example 4.9: The wholesale prices of a commodity for seven consecutive days in a month is as follows:

Days	:	1	2	3	4	5	6	7
Commodity price/quintal	:	240	260	270	245	255	286	264

Calculate the variance and standard deviation.

Solution: The computations for variance and standard deviation are shown in Table 4.7.

Table 4.7 Computations of Variance and Standard Deviation by Actual Mean Method

Observation (x)	$x - \bar{x} = x - 260$	$(x - \bar{x})^2$
240	-20	400
260	0	0
270	10	100
245	-15	225
255	-5	25
286	26	676
264	4	16
1820		1442

$$\bar{x} = \frac{\sum x}{N} = \frac{1820}{7} = 260$$

$$\text{Variance } \sigma^2 = \frac{\sum (x - \bar{x})^2}{N} = \frac{1442}{7} = 206$$

$$\text{Standard deviation } \sigma = \sqrt{\sigma^2} = \sqrt{206} = 14.352$$

In this question, if we take deviation from an assumed A.M. = 255 instead of actual A.M. = 260. The calculations then for standard deviation will be as shown in Table 4.8.

Table 4.8 Computation of Standard Deviation by Assumed Mean Method

Observation (x)	$d = x - A = x - 255$	d^2
240	-15	225
260	5	25
270	15	225
245	-10	100
(255) $\leftarrow A$	0	0
286	31	961
264	9	81
	35	1617

$$\text{Standard deviation } \sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2} = \sqrt{\frac{1617}{7} - \left(\frac{35}{7}\right)^2} \\ = \sqrt{231 - 25} = \sqrt{206} = 14.352$$

This result is same as obtained earlier in Table 4.7.

Remark: When actual A.M. is not a whole number, assumed A.M. method should be used to reduce the computation time.

Example 4.10: A study of 100 engineering companies gives the following information

Profit (Rs in crore) : 0–10	10–20	20–30	30–40	40–50	50–60
Number of companies : 8	12	20	30	20	10

Calculate the standard deviation of the profit earned.

Solution: Let assumed mean, A be 35 and the value of h be 10. Calculations for standard deviation are shown in Table 4.9.

Table 4.9 Calculations of Standard Deviation

Profit (Rs in crore)	Mid-value (m)	$d = \frac{m-A}{h} = \frac{m-35}{10}$	f	fd	fd^2
0–10	5	-3	8	-24	72
10–20	15	-2	12	-24	48
20–30	25	-1	20	-20	20
30–40	(35) $\leftarrow A$	0	30	0	0
40–50	45	1	20	20	20
50–60	55	2	10	20	40
				-28	200

$$\text{Standard deviation, } \sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times h \\ = \sqrt{\frac{200}{100} - \left(\frac{-28}{100}\right)^2} \times 10 = \sqrt{2 - 0.078} \times 10 = 13.863$$

Example 4.11: Mr. Gupta, a retired government servant is considering investing his money in two proposals. He wants to choose the one that has higher average net present value and lower standard deviation. The relevant data are given below. Can you help him in choosing the proposal?

<i>Proposal A:</i>	<i>Net Present Value (NPV)</i>	<i>Chance of the Possible Outcome of NPV</i>
	1559	0.30
	5662	0.40
	9175	0.30
<i>Proposal B:</i>	<i>Net Present Value (NPV)</i>	<i>Chance of the Possible Outcome of NPV</i>
	- 10,050	0.30
	5,812	0.40
	20,584	0.30

Solution: To suggest to Mr. Gupta a proposal for high average net present value, first calculate the expected (average) net present value for both the proposals.

$$\begin{aligned} \text{Proposal A: } \text{Expected NPV} &= 1559 \times 0.30 + 5662 \times 0.40 + 9175 \times 0.30 \\ &= 467.7 + 2264.8 + 2752.5 = \text{Rs } 5485 \end{aligned}$$

$$\begin{aligned} \text{Proposal B: } \text{Expected NPV} &= -10,050 \times 0.30 + 5,812 \times 0.40 + 20,584 \times 0.30 \\ &= -3015 + 2324.8 + 6175.2 = \text{Rs } 5485 \end{aligned}$$

Since the expected NPV in both the cases is same, he would like to choose the less risky proposal. For this we have to calculate the standard deviation in both the cases.

Standard deviation for proposal A:

<i>NPV(x_i)</i>	<i>Expected NPV(̄x)</i>	<i>x - ̄x</i>	<i>f</i>	<i>f(x - ̄x)²</i>
1559	5485	-3926	0.30	46,24,042.8
5662	5485	177	0.40	12,531.6
9175	5485	3690	0.30	40,84,830.0
			1.00	87,21,404.4

$$s_A = \sqrt{\frac{\sum f(x - \bar{x})^2}{\sum f}} = \sqrt{87,21,404.4} = \text{Rs } 2953.20$$

Standard deviation for proposal B:

<i>NPV(x_i)</i>	<i>Expected NPV(̄x)</i>	<i>x - ̄x</i>	<i>f</i>	<i>f(x - ̄x)²</i>
-10,050	5485	-15,535	0.30	7,24,00,867.5
5,812	5485	327	0.40	42,771.6
20,584	5485	15,099	0.30	6,83,93,940
			1.00	14,08,37,579

$$s_B = \sqrt{\frac{\sum f(x - \bar{x})^2}{\sum f}} = \sqrt{14,08,37,579} = \text{Rs } 11,867.50$$

The $s_A < s_B$ indicates uniform net profit for proposal A. Thus proposal A may be chosen.

4.5.3 Mathematical Properties of Standard Deviation

- Combined standard deviation:** The combined standard deviation of two sets of data containing n_1 and n_2 observations with means \bar{x}_1 and \bar{x}_2 and standard deviations σ_1 and σ_2 respectively is given by

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

where

σ_{12} = combined standard deviation

$$d_1 = \bar{x}_{12} - \bar{x}_1 ; \quad d_2 = \bar{x}_{12} - \bar{x}_2$$

and

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \text{ (combined arithmetic mean)}$$

This formula for combined standard deviation of two sets of data can be extended to compute the standard deviation of more than two sets of data on the same lines.

- 2. Standard deviation of natural numbers:** The standard deviation of the first n natural numbers is given by

$$\sigma = \sqrt{\frac{1}{12}(n^2 - 1)}$$

For example, the standard deviation of the first 100 (i.e., from 1 to 100) natural numbers will be

$$\sigma = \sqrt{\frac{1}{12}(100^2 - 1)} = \sqrt{\frac{1}{12}(9999)} = \sqrt{833.25} = 28.86$$

- 3. Standard deviation is independent of change of origin but not of scale.**

Example 4.12: For a group of 50 male workers, the mean and standard deviation of their monthly wages are Rs 6300 and Rs 900 respectively. For a group of 40 female workers, these are Rs 5400 and Rs 600 respectively. Find the standard deviation of monthly wages for the combined group of workers. [Delhi Univ., MBA, 2002]

Solution: Given that $n_1 = 50, \bar{x}_1 = 6300, \sigma_1 = 900$

$$n_2 = 40, \bar{x}_2 = 5400, \sigma_2 = 600$$

$$\text{Then, combined mean, } \bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{50 \times 6300 + 40 \times 5400}{50 + 40} = 5,900$$

and combined standard deviation

$$\begin{aligned} \sigma_{12} &= \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}} \\ &= \sqrt{\frac{50(8,10,000 + 1,60,000) + 40(3,60,000 + 2,50,000)}{50 + 40}} = \text{Rs 900} \end{aligned}$$

$$\text{where } d_1 = \bar{x}_{12} - \bar{x}_1 = 5900 - 6300 = -400$$

$$d_2 = \bar{x}_{12} - \bar{x}_2 = 5900 - 5400 = 500$$

Example 4.13: A study of the age of 100 persons grouped into intervals 20–22, 22–24, 24–26,... revealed the mean age and standard deviation to be 32.02 and 13.18 respectively. While checking, it was discovered that the observation 57 was misread as 27. Calculate the correct mean age and standard deviation. [Delhi Univ., MBA, 1997]

Solution: From the data given in the problem, we have $\bar{x} = 32.02, \sigma = 13.18$ and $N = 100$. We know that

$$\bar{x} = \frac{\sum f x}{N} \text{ or } \sum f x = N \times \bar{x} = 100 \times 32.02 = 3202$$

$$\begin{aligned} \text{and } \sigma^2 &= \frac{\sum f x^2}{N} - (\bar{x})^2 \text{ or } \sum f x^2 = N[\sigma^2 + (\bar{x})^2] = 100[(13.18)^2 + (32.02)^2] \\ &= 100[173.71 + 1025.28] = 100 \times 1198.99 \\ &= 1,19,899 \end{aligned}$$

On re-placing the correct observation, we get

$$\Sigma fx = 3202 - 27 + 57 = 3232.$$

$$\text{Also } \Sigma fx^2 = 1,19,899 - (27)^2 + (57)^2 = 1,19,899 - 729 + 3248 = 1,22,419$$

$$\text{Thus, correct A.M. is } \bar{x} = \frac{\Sigma fx}{N} = \frac{3232}{100} = 32.32.$$

$$\begin{aligned} \text{and correct variance is } \sigma^2 &= \frac{\Sigma fx^2}{N} - (\bar{x})^2 = \frac{1,22,419}{100} - (32.32)^2 \\ &= 1224.19 - 1044.58 = 179.61 \end{aligned}$$

$$\text{or correct standard deviation is, } \sigma = \sqrt{\sigma^2} = \sqrt{179.61} = 13.402.$$

Example 4.14: The mean of 5 observations is 15 and the variance is 9. If two more observations having values -3 and 10 are combined with these 5 observations, what will be the new mean and variance of 7 observations.

Solution: From the data of the problem, we have $\bar{x} = 15$, $s^2 = 9$ and $n = 5$. We know that

$$\bar{x} = \frac{\sum x}{n} \quad \text{or} \quad \sum x = n \times \bar{x} = 5 \times 15 = 75$$

If two more observations having values -3 and 10 are added to the existing 5 observations, then after adding these 6th and 7th observations, we get

$$\sum x = 75 - 3 + 10 = 82$$

$$\text{Thus, the new A.M. is, } \bar{x} = \frac{\sum x}{n} = \frac{82}{7} = 11.71$$

$$\text{Variance, } s^2 = \frac{\sum x^2}{n} - (\bar{x})^2$$

$$9 = \frac{\sum x^2}{n} - (15)^2 \quad \text{or} \quad \sum x^2 = 1170$$

On adding two more observations, i.e., -3 and 10, we get

$$\sum x^2 = 1170 + (-3)^2 + (10)^2 = 1279$$

$$\text{Variance, } s^2 = \frac{\sum x^2}{n} - (\bar{x})^2 = \frac{1279}{7} - (11.71)^2 = 45.59$$

Hence, the new mean and variance of 7 observations is 11.71 and 45.59 respectively.

4.5.4 Chebyshev's Theorem

Standard deviation measures the variation among observations in a set of data. If the standard deviation value is small, then values in the data set cluster close to the mean. Conversely, a large standard deviation value indicates that the values are scattered more widely around the mean. The Russian mathematician P. L. Chebyshev (1821–1894) developed a result called **Chebyshev's theorem** that allows us to determine the proportion of data values that fall within a specified number of standard deviation from the mean value. The theorem states that:

For any set of data (population or sample) and any constant z greater than 1 (but need not be an integer), the proportion of the values that lie within z standard deviations on either side of the mean is at least $\{1 - (1/z^2)\}$. That is

$$\text{RF} [|x - \mu| \leq z \sigma] \geq 1 - \frac{1}{z^2}$$

where RF = relative frequency of a distribution.

$z = \frac{x - \mu}{\sigma} \leftarrow$ population standardized score for an observation x from the population, that is, number of standard deviations a value, x is away from the mean μ (sample or population)

$$= \frac{x - \bar{x}}{s} \leftarrow \text{sample standard score}$$

Chebyshev's theorem: A statement about the proportion of observations that must lie within σ , 2σ , and 3σ deviations from the mean (population or sample distribution).

Chebyshev's theorem states *at least* what percentage of values will fall within z standard deviations in any distribution. However, for a symmetrical, bell-shaped distribution as shown in Fig. 4.4, theorem states *approximately* what percentage of values will fall within z standard deviations.

The relationships involving the mean, standard deviation and the set of observations are called the *empirical rule*, or *normal rule*.

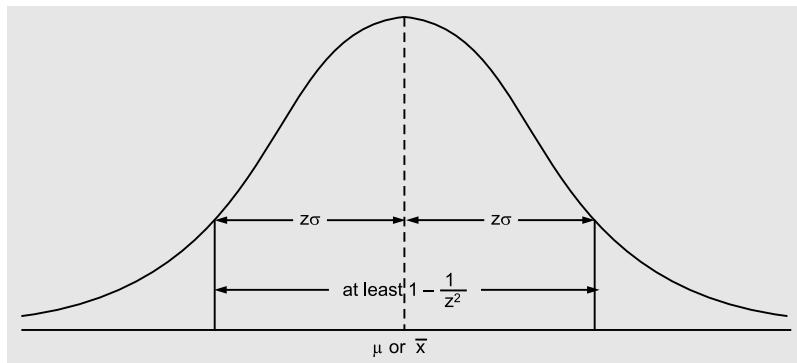


Figure 4.4
Chebyshev's Theorem

Some of the implications of the statement of the theorem with $z = 2, 3$, and 4 standard deviations are as follows:

- (i) The proportion of all x -values in any set of data to fall within the range $\mu \pm 2\sigma$

$$\text{is at least } 1 - \frac{1}{2^2} = \frac{3}{4} = 0.75 \text{ or } 75 \text{ per cent.}$$

That is, at least three of four values or 75 per cent values must lie within ± 2 standard deviations from the mean.

- (ii) The proportion of all x -values in any set of data must lie within the range $\mu \pm 3\sigma$

$$\text{is at least } 1 - \frac{1}{3^2} = \frac{8}{9} = 88.9 \text{ percent.}$$

That is, at least eight of nine values or 88.9 per cent values must lie within ± 3 standard deviations from the mean.

- (iii) The proportion of all x -values in any set of data must lie within the range $\mu \pm 4\sigma$

$$\text{is at least } 1 - \frac{1}{4^2} = \frac{15}{16} = 93.75 \text{ per cent.}$$

This theorem has its own limitation as it emphasizes on the word, 'at least'. For example for $z = 1$, we have, $1 - \frac{1}{1^2} = 0$, which means that the proportion of all x -values to fall within the range $\mu \pm \sigma$ is zero. This result does not give any information.

The theorem is applicable to any data set regardless of the shape of the frequency distribution of values. For example, assume that the marks obtained by 100 students in business statistics had a mean of 70 per cent and standard deviation of 10 per cent. Then number of students who obtained marks between 50 and 85 will be determined as follows:

- (a) For 50 per cent marks, $z = (50 - 70)/10 = -2$ indicates that 50 is 2 standard deviations below the mean,
- (b) For 85 per cent marks, $z = (85 - 70)/10 = 1.5$ indicates that 85 is 1.5 standard deviations above the mean.

Now applying the Chebyshev's theorem with $z = 2.0$, we have

$$\left(1 - \frac{1}{z^2}\right) = \left[1 - \frac{1}{(2.0)^2}\right] = 0.75$$

This indicates that at least 75 per cent of the students must have obtained marks between 50 and 85.

Empirical Rule

For symmetrical, bell-shaped frequency distribution (also called normal curve), the range within which a given percentage of values of the distribution are likely to fall within a specified number of standard deviations of the mean is determined as follows:

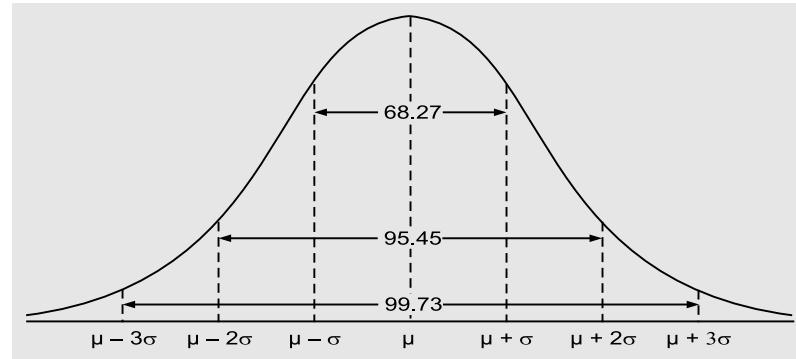
$\mu \pm \sigma$ covers approximately 68.27 per cent of values in the data set

$\mu \pm 2\sigma$ covers approximately 95.45 per cent of values in the data set

$\mu \pm 3\sigma$ covers approximately 99.73 per cent of values in the data set

These ranges are illustrated in Fig. 4.5.

Figure 4.5
Area under Normal Curve



For a symmetrical and bell-shaped distribution, relationships among three measures of variation are given in Table 4.10.

Table 4.10 Relationship Among Measures of Variation

Measures of Variation	Percentage of Values Scatter Around the Mean Value, μ			Size of Measure of Variation to Standard Deviation at
	$\pm \sigma$	$\pm 2\sigma$	$\pm 3\sigma$	
Q.D.	50.00	82.30	95.70	0.6748
MAD	57.50	88.90	98.30	0.7979
S.D.	68.27	95.45	99.73	1.0000

Relationship between Different Measures of Variation

$$(a) \text{Quartile deviation (Q.D.)} = \frac{2}{3}\sigma$$

$$\text{Mean absolute deviation (MAD)} = \frac{4}{5}\sigma$$

$$(b) \text{Quartile deviation} = \frac{5}{6} \text{MAD}$$

$$\text{Standard deviation} = \frac{5}{4} \text{MAD} \quad \text{or} \quad \frac{3}{2} \text{Q.D.}$$

$$(c) \text{Mean absolute deviation} = \frac{6}{5} \text{Q.D.}$$

These relationships are applicable only to symmetrical distributions.

Example 4.15: Suppose you are in charge of rationing in a state affected by food shortage. The following reports arrive from a local investigator:

Daily caloric value of food available per adult during current period:

Area	Mean	Standard Deviation
A	2500	400
B	2000	200

The estimated requirement of an adult is taken as 2800 calories daily and the absolute minimum is 1350. Comment on the reported figures and determine which area in your opinion, need more urgent attention.

Solution: Taking into consideration the entire population of the two areas, we have

$$\text{Area A: } \mu + 3\sigma = 2500 + 3 \times 400 = 3700 \text{ calories}$$

$$\mu - 3\sigma = 2500 - 3 \times 400 = 1300 \text{ calories}$$

This shows that there are adults who are taking even less amount of calories, that is, 1300 calories as compared to the absolute minimum requirement of 1350 calories.

$$\text{Area B: } \mu + 3\sigma = 2000 + 3 \times 200 = 2600 \text{ calories}$$

$$\mu - 3\sigma = 2000 - 3 \times 200 = 1400 \text{ calories}$$

These figures are satisfying the requirement of daily calorific need. Hence, area A needs more urgent attention.

Example 4.16: The following data give the number of passengers travelling by airplane from one city to another in one week.

115 122 129 113 119 124 132 120 110 116

Calculate the mean and standard deviation and determine the percentage of class that lie between (i) $\mu \pm \sigma$, (ii) $\mu \pm 2\sigma$, and (iii) $\mu \pm 3\sigma$. What percentage of cases lie outside these limits?

Solution: The calculations for mean and standard deviation are shown in Table 4.11.

Table 4.11 Calculations of Mean and Standard Deviation

x	$x - \bar{x}$	$(x - \bar{x})^2$
115	-5	25
122	2	4
129	9	81
113	-7	49
119	-1	1
124	4	16
132	12	144
120	0	0
110	-10	100
116	-4	16
1200	0	436

$$\mu = \frac{\sum x}{N} = \frac{1200}{10} = 120 \quad \text{and} \quad \sigma^2 = \frac{\sum (x - \bar{x})^2}{N} = \frac{436}{10} = 43.6$$

Therefore $\sigma = \sqrt{\sigma^2} = \sqrt{43.6} = 6.60$

The percentage of cases that lie between a given limit are as follows:

Interval	Values within Interval	Percentage of Population	Percentage Falling Outside
$\mu \pm \sigma = 120 \pm 6.60$ = 113.4 and 126.6	113, 115, 116, 119 120, 122, 124	70%	30%
$\mu \pm 2\sigma = 120 \pm 2(6.60)$ = 106.80 and 133.20	110, 113, 115, 116, 119 120, 122, 124, 129, 132	100%	nil

Example 4.17: A collar manufacturer is considering the production of a new collar to attract young men. Thus following statistics of neck circumference are available based on measurement of a typical group of the college students:

Mid value (in inches) :	12.0	12.5	13.0	13.5	14.0	14.5	15.0	15.5	16.0
Number of students :	2	16	36	60	76	37	18	3	2

Compute the standard deviation and use the criterion $\bar{x} \pm 3\sigma$, where σ is the standard deviation and \bar{x} is the arithmetic mean to determine the largest and smallest size of the collar he should make in order to meet the needs of practically all the customers bearing in mind that collar are worn on average half inch longer than neck size.

Solution: Calculations for mean and standard deviation in order to determine the range of collar size to meet the needs of customers are shown in Table 4.12.

Table 4.12 Calculations for Mean and Standard Deviation

Mid-value (in inches)	Number of students	$\frac{x - A}{h} = \frac{x - 14}{0.5}$	fd	fd^2
12.0	2	-4	-8	32
12.5	16	-3	-48	144
13.0	36	-2	-72	144
13.5	60	-1	-60	60
(14.0) ← A	76	0	0	0
14.5	37	1	37	37
15.0	18	2	36	72
15.5	3	3	9	27
16.0	2	4	8	32
$N = 250$			-98	548

$$\text{Mean}, \bar{x} = A + \frac{\sum fd}{N} \times h = 14.0 - \frac{98}{250} \times 0.5 = 14.0 - 0.195 = 13.805$$

$$\begin{aligned} \text{Standard deviation}, \sigma &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times h = \sqrt{\frac{548}{250} - \left(\frac{-98}{250}\right)^2} \times 0.5 \\ &= \sqrt{2.192 - 0.153} \times 0.5 = 1.427 \times 0.5 = 0.7135 \end{aligned}$$

$$\begin{aligned} \text{Largest and smallest neck size} &= \bar{x} \pm 3\sigma = 13.805 \pm 3 \times 0.173 \\ &= 11.666 \text{ and } 15.944. \end{aligned}$$

Since all the customers are to wear collar half inch longer than their neck size, 0.5 is to be added to the neck size range given above. The new range then becomes:

(11.666 + 0.5) and (15.944 + 0.5) or 12.165 and 16.444, i.e. 12.2 and 16.4 inches.

Example 4.18: The breaking strength of 80 'test pieces' of a certain alloy is given in the following table, the unit being given to the nearest thousand grams per square inch;

Breaking Strength	Number of Pieces
44–46	3
46–48	24
48–50	27
50–52	21
52–54	5

Calculate the average breaking strength of the alloy and the standard deviation. Calculate the percentage of observations lying between $\bar{x} \pm 2\sigma$.

[Vikram Univ., MBA, 2000]

Solution: The calculations for mean and standard deviation are shown in Table 4.13.

Table 4.13 Calculations for Mean and Standard Deviation

<i>Breaking Strength</i>	<i>Number of Pieces(f)</i>	<i>Mid-value (m)</i>	$d = (m - A)/h$ $= (m - 49)/2$	fd	fd^2
44–46	3	45	-2	-6	12
46–48	24	47	-1	-24	24
48–50	27	A → 49	0	0	0
50–52	21	51	1	21	21
52–54	5	53	2	10	20
	80			1	77

$$\text{Mean, } \bar{x} = A + \frac{\sum fd}{N} \times h = 49 + \frac{1}{80} \times 2 = 49.025$$

$$\begin{aligned} \text{Standard deviation, } \sigma &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times h = \sqrt{\frac{77}{80} - \left(\frac{1}{80}\right)^2} \times 2 \\ &= \sqrt{0.9625 - 0.000} \times 2 = 0.9810 \times 2 = 1.962 \end{aligned}$$

Breaking strength of pieces in the range, $\bar{x} \pm 2\sigma$ is

$$\begin{aligned} \bar{x} \pm 2\sigma &= 49.025 \pm 2 \times 1.962 \\ &= 45.103 \text{ and } 52.949 = 45 \text{ and } 53 \text{ (approx.)} \end{aligned}$$

To calculate the percentage of observations lying between $\bar{x} \pm 2\sigma$, we assume that the number of observations (pieces) are equally spread within lower and upper boundary of each class interval (breaking strength). Since 45 is the mid-point of the class interval 44–46 with the frequency 3, therefore there are 1.5 frequencies at 45. Similarly, at 53 the frequency would be 2.5. Hence the total number of observations (frequencies) between 45 and 53 are $= 1.5 + 24 + 27 + 21 + 2.5 = 76$. So the percentage of observations lying within $\bar{x} \pm 2\sigma$ would be $(76/80) \times 100 = 95$ per cent.

4.5.5 Coefficient of Variation

Standard deviation is an absolute measure of variation and expresses variation in the same unit of measurement as the arithmetic mean or the original data. A relative measure called the **coefficient of variation** (CV), developed by Karl Pearson is very useful measure for (i) comparing two or more data sets expressed in different units of measurement (ii) comparing data sets that are in same unit of measurement but the mean values of data sets in a comparable field are widely dissimilar (such as mean wages received per month by the top management personnel and labour class personnel of a large organization).

Thus, in view of this limitation we need to convert absolute measure of variation, that is, S.D. into a relative measure, which can be helpful in comparing the variability of two or more sets of data. The new measure, coefficient of variation (CV), measures the standard deviation relative to the mean in percentages. In other words, CV indicates how large the standard deviation is in relation to the mean and is computed as follows:

$$\text{Coefficient of variation (CV)} = \frac{\text{Standard deviation}}{\text{Mean}} \times 100 = \frac{\sigma}{\bar{x}} \times 100$$

Multiplying by 100 converts the decimal to a percent.

The set of data for which the coefficient of variation is low is said to be more uniform (consistent) or more homogeneous (stable).

Coefficient of variation: A measure of relative variability computed by dividing the standard deviation by the mean, then multiplying by 100.

Example 4.19: The weekly sales of two products A and B were recorded as given below:

Product A :	59	75	27	63	27	28	56
Product B :	150	200	125	310	330	250	225

Find out which of the two shows greater fluctuation in sales.

Solution: For comparing the fluctuation in sales of two products, we will prefer to calculate coefficient of variation for both the products.

Product A: Let $A = 56$ be the assumed mean of sales for product A.

Table 4.14 Calculations of the Mean and Standard Deviation

Sales (x)	Frequency (f)	$d = x - A$	fd	fd^2
27	2	-29	-58	1682
28	1	-28	-28	784
(56) $\leftarrow A$	1	0	0	0
59	1	3	3	9
63	1	7	7	49
75	1	19	19	361
		<u>7</u>	<u>-57</u>	<u>2885</u>

$$\bar{x} = A + \frac{\sum fd}{\sum f} = 56 - \frac{57}{7} = 47.86$$

$$\begin{aligned}s_A^2 &= \frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f} \right)^2 = \frac{2885}{7} - \left(-\frac{57}{7} \right)^2 \\ &= 412.14 - 66.30 = 345.84\end{aligned}$$

$$s_A = \sqrt{345.84} = 18.59$$

$$\text{Then } CV(A) = \frac{s_A}{\bar{x}} \times 100 = \frac{18.59}{47.86} \times 100 = 38.84 \text{ per cent}$$

Product B: Let $A = 225$ be the assumed mean of sales for product B.

Table 4.15 Calculations of Mean and Standard Deviation

Sales (x)	Frequency (f)	$d = x - A$	fd	fd^2
125	1	-100	-100	10,000
150	1	-75	-75	5625
200	1	-25	-25	625
225	1	0	0	0
250	1	25	25	625
310	1	85	85	7225
330	1	105	105	11,025
		<u>7</u>	<u>15</u>	<u>35,125</u>

$$\bar{x} = A + \frac{\sum fd}{\sum f} = 225 + \frac{15}{7} = 227.14$$

$$\begin{aligned}s_B^2 &= \frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f} \right)^2 = \frac{35,125}{7} - \left(\frac{15}{7} \right)^2 \\ &= 5017.85 - 4.59 = 5013.26\end{aligned}$$

$$\text{or } s_B = \sqrt{5013.26} = 70.80$$

$$\text{Then } CV(B) = \frac{s_B}{\bar{x}} \times 100 = \frac{70.80}{227.14} \times 100 = 31.17 \text{ per cent}$$

Since the coefficient variation for product A is more than that of product B, therefore the sales fluctuation in case of product A is higher.

Example 4.20: From the analysis of monthly wages paid to employees in two service organizations X and Y, the following results were obtained:

	<i>Organization X</i>	<i>Organization Y</i>
Number of wage-earners	550	650
Average monthly wages	5000	4500
Variance of the distribution of wages	900	1600

- (a) Which organization pays a larger amount as monthly wages?
 (b) In which organization is there greater variability in individual wages of all the wage earners taken together?

Solution: (a) For finding out which organization X or Y pays larger amount of monthly wages, we have to compare the total wages:

Total wage bill paid monthly by X and Y is

$$X : n_1 \times \bar{x}_1 = 550 \times 5000 = \text{Rs. } 27,50,000$$

$$Y : n_2 \times \bar{x}_2 = 650 \times 4500 = \text{Rs. } 29,25,000$$

Organization Y pays a larger amount as monthly wages as compared to organization X.

(b) For calculating the combined variation, we will first calculate the combined mean as follows:

$$\begin{aligned} \bar{x}_{12} &= \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{27,50,000 + 29,25,000}{1200} = \text{Rs } 4729.166 \\ \sigma_{12} &= \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}} \\ &= \sqrt{\frac{550(900 + 73,351.05) + 650(1600 + 52,517.05)}{550 + 650}} \\ &= \sqrt{\frac{4,08,38,080.55 + 3,51,76,082.50}{1200}} = \sqrt{63345.13} = 251.68 \end{aligned}$$

where $d_1 = \bar{x}_{12} - \bar{x}_1 = 4729.166 - 5000 = -270.834$

$d_2 = \bar{x}_{12} - \bar{x}_2 = 4729.166 - 4500 = 229.166$

Conceptual Questions 4B

8. What purpose does a measure of variation serve? In the light of these, comment on some of the well-known measures of variation.
9. What do you understand by 'coefficient of variation'? Discuss its importance in business problems.
[UP Tech. Univ., MBA, 2000]
10. When is the variance equal to the standard deviation? Under what circumstances can variance be less than the standard deviation? Explain.
11. (a) Explain and illustrate how the measures of variation afford a supplement to the information about frequency distribution furnished by averages.
[Delhi Univ., MBA, 2001]
 (b) Describe various methods of measuring variation. Which of these do you consider as the best and why?
12. Explain the advantages of standard deviation as a measure of variation over range and the average deviation. Under what circumstances will the variance of a variable be zero?
13. Comment on the comparative merits and demerits of measures of variation.
14. Explain the term 'variation'. What purpose does a measure of variation serve? In the light of these, comment on some of the well-known measures of variation.
[Delhi Univ., MBA, 1998]
15. Describe the various methods of measuring variation along with their respective merits and demerits.
[Delhi Univ., MBA, 1998]
16. It has been said that the lesser the variability that exists, the more an average is representative of a set of data. Comment.
17. (a) What information is provided by variance or standard deviation?
 (b) What additional information about a set of data is provided by a measure of variability that is not obtained from an average?
18. What advantages are associated with variance and standard deviation relative to range as the measure of variability?
19. Suppose you read a published statement that the average amount of food consumption in this country is adequate; the overall conclusion based upon the statement is that

everyone is properly fed. Criticize the conclusion in terms of the concept of variability as it relates to the use of averages. [Delhi Univ., MBA, 2000]

- 20.** The Vice-President, Sales has been studying records regarding the performance of his sales representatives. He has noticed that in the last 2 years, the average level of sales per representative has remained the same, while the distribution of the sales levels has widened. The sales levels from this period have significantly larger variations from the mean than in any of the previous 2 year periods for which he has records. What conclusions might be drawn from these observations? [Delhi Univ., MBA, 1999]
- 21.** Explain Chebyshev's theorem which provides an approximation to the spread of a set of observations on either side of the mean.

- 22.** Two economists are studying fluctuations in the price of gold. One is examining the period of 1998–2002. The other is examining the period of 1995–1999. What differences would you expect to find in the variability of their data?
- 23.** How would you reply to the following statement: 'Variability is not an important factor because even though the outcome is more uncertain, you still have an equal chance of falling either above or below the median. Therefore on an average, the outcome will be the same.'
- 24.** A retailer uses two different formulas for predicting monthly sales. The first formula has an average miss of 700 records, and a standard deviation of 35 records. The second formula has an average miss of 300 records, and a standard deviation of 16. Which formula is relatively less accurate?

Self-Practice Problems 4B

- 4.9** Find the average deviation from mean for the following distribution:

Quantity demanded (in units) :

60	61	62	63	64	65	66	67	68
----	----	----	----	----	----	----	----	----

Frequency :

2	0	15	29	25	12	10	4	3
---	---	----	----	----	----	----	---	---

- 4.10** Find the average deviation from mean for the following distribution:

Dividend yield :

0–3	3–6	6–9	9–12	12–15	15–18	18–21
-----	-----	-----	------	-------	-------	-------

Number of companies :

2	7	10	12	9	6	4
---	---	----	----	---	---	---

- 4.11** Find the average deviation from median for the following distribution:

Sales (Rs '000) :

1–3	3–5	5–7	7–9	9–11	11–13	13–15	15–17
-----	-----	-----	-----	------	-------	-------	-------

Number of shops :

6	53	85	56	21	26	4	4
---	----	----	----	----	----	---	---

- 4.12** In a survey of 48 engineering companies, following data was collected:

Level of profit (Rs in lakh) : 10 11 12 13 14

Number of companies : 3 12 18 12 3

Calculate the variance and standard deviation for the distribution.

- 4.13** A manufacturer of T-shirts approaches you with the following information

Length of shoulder (in inches) :

12.0	12.5	13.0	13.5	14	14.5	15	15.5	16
------	------	------	------	----	------	----	------	----

Frequency:

5	20	30	43	60	56	37	16	3
---	----	----	----	----	----	----	----	---

Calculate the standard deviation and advise the manufacturer as to the largest and the smallest shoulder size T-shirts he should make in order to meet the needs of his customers.

- 4.14** A charitable organization decided to give old-age pension to people over sixty years of age. The scales of pension were fixed as follows:

Age Group	Pension/month (Rs)
60–65	200
65–70	250
70–75	300
75–80	350
80–85	400

The ages of 25 persons who secured the pension are as given below:

74	62	84	72	61	83	72	81	64
71	63	61	60	67	74	64	79	73
75	76	69	68	78	66	67		

Calculate the monthly average pension payable per person and the standard deviation.

- 4.15** Two automatic filling machines A and B are used to fill tea in 500 g cartons. A random sample of 100 cartons on each machine showed the following:

Tea Contents (in g)	Machine A	Machine B
485–490	12	10
490–495	18	15
495–500	20	24
500–505	22	20
505–510	24	18
510–515	4	13

Comment on the performance of the two machines on the basis of average filling and dispersion.

- 4.16** An analysis of production rejects resulted in the following observations

No. of Rejects per Operator	No. of Operators	No. of Rejects per Operator	No. of Operators
21–25	5	41–45	15
26–30	15	46–50	12
31–35	28	51–55	3
36–40	42		

Calculate the mean and standard deviation.

[Delhi Univ., MBA, 2000]

- 4.17** Blood serum cholesterol levels of 10 persons are as under:

240 260 290 245 255 288 272 263 277 250

Calculate the standard deviation with the help of assumed mean

- 4.18** 32 trials of a process to finish a certain job revealed the following information:

Mean time taken to complete the job = 80 minutes

Standard deviation = 16 minutes

Another set of 8 trials gave mean time as 100 minutes and standard deviation equalled to 25 minutes.

Find the combined mean and standard deviation.

- 4.19** From the analysis of monthly wages paid to workers in two organizations X and Y, the following results were obtained:

	X	Y
Number of wage-earners	: 550	600
Average monthly wages (Rs)	: 1260	1348.5
Variance of distribution of wages (Rs)	: 100	841

Obtain the average wages and the variability in individual wages of all the workers in the two organizations taken together.

- 4.20** An analysis of the results of a budget survey of 150 families showed an average monthly expenditure of Rs 120 on food items with a standard deviation of Rs 15. After the analysis was completed it was noted that the figure recorded for one household was wrongly taken as Rs 15 instead of Rs 105. Determine the correct value of the average expenditure and its standard deviation.

- 4.21** The standard deviation of a distribution of 100 values was Rs 2. If the sum of the squares of the actual values was Rs 3,600, what was the mean of this distribution?

- 4.22** An air-charter company has been requested to quote a realistic turn-round time for a contract to handle certain imports and exports of a fragile nature.

The contract manager has provided the management accountant with the following analysis of turn-round times for similar goods over a given twelve-monthly period.

Turn-round Time (in hours)	Frequency
Less than 2	25
2 and < 4	36
4 and < 6	66
6 and < 8	47
8 and < 10	26
10 and < 12	18
12 and < 14	2

(a) Calculate mean and standard deviation.

(b) Advise the contract manager about the turn-round time to be quoted using

- (i) mean plus one standard deviation;
- (ii) mean plus two standard deviations.

- 4.23** The following relationship holds between two measures of temperature:

$$F = 32 + \frac{9}{5} C$$

where F and C denote the degree in daily average temperature measured in Fahrenheit and Centigrade.

If the variance of daily average temperature in a city throughout the year is 25°C, what is the variance in °F for that year and vice-versa.

- 4.24** The hourly output of a new machine is four times that of the old machine. If the variance of the hourly output of the old machine in a period of n hours is 16, what is the variance of the hourly output of the new machine in the same period of n hours.

- 4.25** The number of cheques cashed each day at the five branches of a bank during the past month has the following frequency distribution:

Number of Cheques	Frequency
0–199	10
200–399	13
400–599	17
600–799	42
800–999	18

The General manager, operations, for the bank, knows that a standard deviation in cheque cashing of more than 200 checks per day creates staffing problem at the branches because of the uneven workload. Should the manager worry about staffing next month?

- 4.26** Mr. Gupta, owner of a bakery, said that the average weekly production level of his company was 11,398 loaves, and the variance was 49,729. If data used to compute the results were collected for 32 weeks, during how many weeks was the production level below 11,175? and Above 11,844?

Coefficient of Variance

- 4.27** Two salesmen selling the same product show the following results over a long period of time:

	Salesman X	Salesman Y
Average sales volume per month (Rs)	30,000	35,000
Standard deviation	2,500	3,600

Which salesman seems to be more consistent in the volume of sales?

- 4.28** Suppose that samples of polythene bags from two manufacturers A and B are tested by a buyer for bursting pressure, giving the following results:

Bursting Pressure	Number of Bags	
	A	B
5.0–9.9	2	9
10.0–14.9	9	11
15.0–19.9	29	18
20.0–24.9	54	32
25.0–29.9	11	27
30.0–34.9	5	13

- (a) Which set of bags has the highest bursting pressure?
 (b) Which has more uniform pressure? If prices are the same, which manufacturer's bags would be preferred by the buyer? Why?

[Delhi Univ., MBA 1997]

- 4.29** The number of employees, average daily wages per employee, and the variance of daily wages per employee for two factories are given below:

	Factory A	Factory B
Number of employees	: 50	100
Average daily wages (Rs)	: 120	85
Variance of daily wages (Rs)	: 9	16

- (a) In which factory is there greater variation in the distribution of daily wages per employee?
 (b) Suppose in factory B, the wages of an employee were wrongly noted as Rs 120 instead of Rs 100. What would be the correct variance for factory B?
4.30 The share prices of a company in Mumbai and Kolkata markets during the last ten months are recorded below:

Month	Mumbai	Kolkata
January	105	108
February	120	117
March	115	120
April	118	130
May	130	100
June	127	125
July	109	125
August	110	120
September	104	110
October	112	135

Determine the arithmetic mean and standard deviation of prices of shares. In which market are the share prices more stable? [HP Univ., MBA 2002]

- 4.31** A person owns two petrol filling stations A and B. At station A, a representative sample of 200 consumers who purchase petrol was taken. The results were as follows:

Number of Litres of Petrol Purchased	Number of Consumers
0 and < 2	15
2 and < 4	40
4 and < 6	65
6 and < 8	40
8 and < 10	30
10 and over	10

A similar sample at station B users showed a mean of 4 litres with a standard deviation of 2.2 litres. At which station is the purchase of petrol relatively more variable?

Hints and Answers

- 4.9** $MAD = 1.239$; $\bar{x} = 63.89$
4.10 $\bar{x} = 10.68$; $MAD = 3.823$
4.11 $Med = 6.612$; $MAD = 2.252$
4.12 $\sigma^2 = 1$ and $\sigma = 1$
4.13 $\bar{x} = 14.013$ inches; $\sigma = 0.8706$ inches; $\bar{x} + 3\sigma = 14.884$ (largest size); $\bar{x} - 3\sigma = 13.142$ (smallest size)
4.14 $\bar{x} = \text{Rs } 280.2$; $\sigma = \text{Rs } 60.765$
4.15 Machine A: $\bar{x}_1 = 499.5$; $\sigma_1 = 7.14$; Machine B: $\bar{x}_2 = 500.5$; $\sigma_2 = 7.40$
4.16 $\bar{x} = 36.96$; $\sigma = 6.375$
4.17 $\sigma = 16.48$
4.18 $\bar{x}_{12} = 84$ minutes; $\sigma_{12} = 19.84$
4.19 $\bar{x}_{12} = \text{Rs } 1306$; $\sigma_{12} = \text{Rs } 53.14$
4.20 Corrected $\bar{x} = \text{Rs } 120.6$ and corrected $\sigma = \text{Rs } 12.4$
4.21 $\bar{x} = 5.66$
4.22 (a) (i) $\bar{x} = 5.68$; (ii) $\sigma = 2.88$
 (b) (i) $\bar{x} + \sigma = 5.68 + 2.88 = 8.56$ hours
 The chance of this turn-round time cover approx. 84%
- (ii) $\bar{x} + 2\sigma = 5.68 + 2(2.88) = 11.44$ hours
 The chance of this turn round time cover approx. 97.7%
4.24 Variance (new machine) = 256 hours
4.25 Mean $\mu = \frac{\sum f x}{N} = \frac{59,000}{100} = 590$ cheques per day
 Standard deviatio, $\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{N}} = \sqrt{\frac{58,70,000}{100}} = 242.48$ cheques per day
 Since standard deviation σ value is more than 200, the manager should worry.
4.26 The standard deviation for the distribution is $\sigma = \sqrt{\sigma^2} = \sqrt{49,729} = 223$. A production of 11,175 loaves is one standard deviation below the mean $(11,398 - 11,175) = 223$. Assuming that the distribution is symmetrical, we know that within $\mu \pm \sigma$ per cent about 68% of all observations fall. The interval from the mean to one standard deviation below the mean would contain about 34 per cent (68 per cent $\div 2$) of the data. Therefore, (50 -

34) = 16 per cent (or approx 5 weeks) of the data would be below 11,175 loaves.

4.27 Salesman X

4.28 Manufacture A: $\bar{x}_1 = 21$, $\sigma_1 = 4.875$ and C.V. = 23.32%

Manufacturer B : $\bar{x}_2 = 21.81$, $\sigma_2 = 7.074$ and C.V. = 32.44%; (a) Bags of manufacturer B have higher bursting pressure; (b) Bags of manufacturer A have more uniform pressure; (c) Bags of manufacturer A should be preferred by buyer as they have uniform pressure.

4.29 (a) CV(A) = 2.5 ;

CV(B) = 4.7. Variation in the distribution of daily wages per employee in factory B is more.

(b) Correct $\Sigma x = 100 \times 85 - 120 + 100 = 8,480$

Correct mean $\bar{x} = 8480/100 = 84.8$

Since $\sigma^2 = (\Sigma x^2/N) - (\bar{x})^2$

or $16 = (\Sigma x^2/100) - (85)^2$

$$= \Sigma x^2 - 7,22,500$$

or $\Sigma x^2 = 7,24,100$

$$\text{Correct } \Sigma x^2 = 7,24,100 - (120)^2 + (100)^2 \\ = 7,19,700$$

$$\text{Correct } \sigma^2 = (7,19,700/100) - (84.8)^2 = 5.96$$

4.30 CV(Mumbai) = 7.24%; CV (Kolkata) = 8.48%. This shows more stability in Mumbai stock market.

4.31 CV(A) = 46.02%; CV(B) = 55%. The purchase of petrol is relatively more variable at station B.

Formulae Used

1. Range, R

Value of highest observation – Value of lowest observation = H – L

$$\text{Coefficient of range} = \frac{H-L}{H+L}$$

2. Interquartile range = $Q_3 - Q_1$

$$\text{Quartile deviation, QD} = \frac{Q_3 - Q_1}{2}$$

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

3. Mean average deviation

For ungrouped data

$$(i) \text{ MAD} = \frac{\sum |x - \bar{x}|}{n}, \text{ for sample}$$

$$(ii) \text{ MAD} = \frac{\sum |x - \mu|}{N}, \text{ for population}$$

$$(iii) \text{ MAD} = \frac{\sum |x - Me|}{n}, \text{ from median}$$

$$\text{For grouped data} \quad \text{MAD} = \frac{\sum f|x - \bar{x}|}{\sum f}$$

4. Coefficient of

$$\text{MAD} = \frac{\text{MAD}}{\bar{x} \text{ or } Me} \times 100$$

5. Variance

Ungrouped data

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{N} = \frac{\sum x^2}{N} - \left(\frac{\sum x}{N} \right)^2$$

$$= \frac{\sum d^2}{N} - \left(\frac{\sum d}{N} \right)^2$$

where $d = x - A$; A is any assumed A.M. value

$$\text{Grouped data, } \sigma^2 = \left[\frac{\sum f d^2}{N} - \left(\frac{\sum f d}{N} \right)^2 \right] h$$

where $d = (m - A)/h$; h is the class interval and m is the mid-value of class intervals.

6. Standard deviation

$$\text{Ungrouped data } \sigma = \sqrt{\sigma^2}$$

$$\text{Grouped data } \sigma = \sqrt{\frac{\sum f d^2}{N} - \left(\frac{\sum f d}{N} \right)^2} \times h$$

$$7. \text{ Coefficient of variation (CV)} = \frac{\sigma}{\bar{x}} \times 100$$

Review Self-Practice Problems

4.32 A petrol filling station has recorded the following data for litres of petrol sold per automobile in a sample of 680 automobiles:

Petrol Sold (Litres)	Frequency
0– 4	74
5– 9	192
10–14	280
15–19	105
20–24	23
25–29	6

Compute the mean and standard deviation for the data.

4.33 A frequency distribution for the duration of 20 long-distance telephone calls in minutes is as follows:

Call Duration (Minutes)	Frequency
4– 7	4
8–11	5
12–15	7
16–19	2
20–23	1
24–27	1

Compute the mean, variance, and standard deviation.

- 4.34** Automobiles travelling on a highway are checked for speed by the police. Following is a frequency distribution of speeds:

Speed (km per hours)	Frequency
45–49	10
50–54	40
55–59	150
60–64	175
65–69	75
70–74	15
75–79	10

What is the mean, variance, and standard deviation of speed for the automobiles travelling on the highway?

- 4.35** A work-standards expert observes the amount of time (in minutes) required to prepare a sample of 10 business letters in the office with observations in ascending order: 5, 5, 5, 7, 9, 14, 15, 15, 16, 18.

- (a) Determine the range and middle 70 per cent range for the sample.
- (b) If the sample mean of the data is 10.9, then calculate the mean absolute deviation and variance.

- 4.36** ABC Stereos, a wholesaler, was contemplating becoming the supplier to three retailers, but inventory shortages have forced him to select only one. ABC's credit manager is evaluating the credit record of these three retailers. Over the past 5 years these retailers' accounts receivable have been outstanding for the following average number of days. The credit manager feels that consistency, in addition to lowest average, is important. Based on relative dispersion, which retailer would make the best customer?

Lee :	62.20	61.80	63.40	63.00	61.70
Forest :	62.50	61.90	63.80	63.00	61.70
Davis :	62.00	61.90	63.00	63.90	61.50

[Delhi Univ., MBA, 1999]

- 4.37** A purchasing agent obtained samples of 60 watt bulbs from two companies. He had the samples tested in his own laboratory for length of life with the following results:

Length of Life (in hours)	Samples from	
	Company A	Company B
1700–1900	10	3
1900–2100	16	40
2100–2300	20	12
2300–2500	8	3
2500–2700	6	2

- (a) Which company's bulbs do you think are better in terms of average life?
- (b) If prices of both the companies are same, which company's bulbs would you buy and why?

[Delhi Univ., MBA, 2000]

- 4.38** The Chief Medical Officer of a hospital conducted a survey of the number of days 200 randomly chosen patients stayed in the hospital following an operation. The data are given below

Hospital stay (in days) :

1–3 4–6 7–9 10–12 13–15 16–18 19–21 22–24

Number of patients:

18 90 44 21 9 9 4 5

- (a) Calculate the mean number of days patients stay in the hospital along with standard deviation of the same.
- (b) How many patients are expected to stay between 0 and 17 days.

- 4.39** A nursing home is well-known in effective use of pain killing drugs for seriously ill patients. In order to know approximately how many nursing staff to employ, the nursing home has begun keeping track of the number of patients that come every week for checkup. Each week the CMO records the number of seriously ill patients and the number of routine patients. The data for the last 5 weeks is as follows:

Seriously ill patients : 33 50 22 27 48

Routine patients : 34 31 37 36 27

- (a) Find the limits within which the middle 75 per cent of seriously ill patients per week should fall.
- (b) Find the limits within which the middle 68 per cent of routine patients per week should fall.

- 4.40** There are a number of possible measures of sales performance, including how consistent a sales person is, in meeting established sales goals. The following data represent the percentage of goal met by each of three sales persons over the last five years

Raman :	88	68	89	92	103
Sindhu :	76	88	90	86	79
Prasad :	104	88	118	88	123

Which salesman is most consistent. Suggest an alternative measure of consistency (if possible).

- 4.41** Gupta Machine Company has a contract with one of its customers to supply machined pump gears. One requirement is that the diameter of its gears be within specific limits. The following data is of diameters (in inches) of a sample of 20 gears:

4.01 4.00 4.02 4.03 4.00 3.98 3.99 3.99

4.01 4.02 3.99 3.98 3.97 4.00 4.02 4.01

4.02 4.00 4.01 3.99

What can Gupta say to his customers about the diameters of 95 per cent of the gears they are receiving?

[Delhi Univ., MBA, 1998]

- 4.42** A production department uses a sampling procedure to test the quality of newly produced items. The department employs the following decision rule at an inspection station: If a sample of 14 items has a variance of more than 0.005, the production line must be shut down for repairs. Suppose the following data have just been collected:

3.43 3.45 3.43 3.48 3.52 3.50 3.39

3.48 3.41 3.38 3.49 3.45 3.51 3.50

Should the production line be shut down? Why or why not?

- 4.43** Police records show the following numbers of daily crime reports for a sample number of days during the winter

months and a sample number of days during the summer months.

Winter : 18 20 15 16 21 20 12 16 19 20
Summer : 28 18 24 32 18 29 23 38 28 18

- Compute the range and inter-quartile range for each period.
- Compute the variance and standard deviation for each period.
- Compute the coefficient of variation for each period.

- 4.44** Public transportation and the automobiles are two options an employee can use to get to work each day. Samples of time (in minutes) recorded for each option are shown below:

Public transportation :

28 29 32 37 33 25 29 32 41 34

Automobile :

29 31 33 32 34 30 31 32 35 33

- Compute the sample mean time to get to work for each option.
- Compute the sample standard deviation for each option.
- On the basis of your results from parts (a) and (b), which method of transportation should be preferred? Explain.

- 4.45** The mean and standard deviation of a set of 100 observations were worked out as 40 and 5 respectively by a computer which, by mistake, took the value 50 in place of 40 for one observation. Find the correct mean and variance. [Lucknow Univ., MBA, 1989]

- 4.46** The number of employees, wages per employee and the variance of the wages per employee for two factories is given below:

	Factory A	Factory B
Number of employees	: 100	150
Average wage per employee per month (Rs)	: 3200	2800
Variance of the wages per employee per month (Rs)	: 625	729

- In which factory is there greater variation in the distribution of wages per employee?
- Suppose in factory B, the wages of an employee were wrongly noted as Rs 3050 instead of Rs 3650, what would be the correct variance for factory B?

[Kumaun Univ., MBA, 1998]

- 4.47** In two factories A and B engaged in the same industry, the average weekly wages and standard deviations are as follows:

Factory	Average Weekly Wages (Rs)	S.D. of Wages (Rs)	No. of Wage Earners
A	460	50	100
B	490	40	80

- Which factory, A or B, pays a higher amount as weekly wages?
- Which factory shows greater variability in the distribution of wages?

- What is the mean and standard deviation of all the workers in two factories taken together?

[HP Univ., MBA ; Vikram Univ., MBA, 1997]

- 4.48** The mean of 5 observations is 4.4 and the variance is 8.24. If three of the five observations are 1, 2 and 6, find the other two.

- 4.49** The mean and standard deviation of normal distribution are 60 and 5 respectively. Find the inter-quartile range and the mean deviation of the distribution:

[Delhi Univ., BCom (H), 1997]

- 4.50** Mean and standard deviation of the following continuous series are 31 and 5.9 respectively. The distribution after taking step deviations is as follows:

Step deviation, d : -3 -2 -1 0 1 2 3

Frequency, f : 10 15 25 25 10 10 5

Determine the actual class intervals.

[Delhi Univ., BCom (H) 1998]

- 4.51** The value of the arithmetic mean and standard deviation of the following frequency distribution of a continuous variable derived from the use of working origin and scale are Rs. 107 and 13.1 respectively. Determine the actual classes.

Step deviation, d : -3 -2 -1 0 +1 +2

Frequency, f : 1 3 4 7 3 2

[Ranchi Univ., MBA, 1998]

- 4.52** The mean and standard deviation of a set of 100 observations were found to be 40 and 5 respectively. But by mistake a value 50 was taken in place of 40 for one observation. Re-calculate the correct mean and standard deviation. [Lucknow Univ., MBA, 1999]

- 4.53** The mean and the standard deviation of a sample of 10 sizes were found to be 9.5 and 2.5 respectively. Later on, an additional observation became available. This was 15.0 and was included in the original sample. Find the mean and the standard deviation of 11 observations.

- 4.54** The Shareholders Research Centre of India has recently conducted a research-study on price behaviour of three leading industrial shares, A, B, and C for the period 1979 to 1985, the results of which are published as follows in its Quarterly Journal:

Share	Average Price (Rs)	Standard Deviation	Current Selling Price (Rs)
A	18.2	5.4	36.00
B	22.5	4.5	34.75
C	24.0	6.0	39.00

- Which share, in your opinion, appears to be more stable in value?

- If you are the holder of all the three shares, which one would you like to dispose of at present, and why? [HP Univ., MCom; Jammu Univ., MCom, 1997]

- 4.55** Find the missing information from the following:

	Group I	Group II	Group III	Combined
Number	50	?	90	200
Std. dev.	6	7	?	7.746
Mean	113	?	115	116

[HP Univ., MBA; Osmania Univ., MBA, 1997]

- 4.56** An analysis of the weekly wages paid to workers in two firms A and B belonging to the same industry, gives the following results:

	Firm A	Firm B
Number of wage-earners	550	650
Average daily wages	50	45
Standard deviation of the distribution of wages	$\sqrt{90}$	$\sqrt{120}$

- (a) Which firm, A or B, pays out a larger amount as daily wages?
 (b) In which firm, A or B, is there greater variability in individual wages?
 (c) What are the measures of (i) average daily wages and (ii) standard deviation in the distribution of individual wages of all workers in the two firms taken together?

[M.D. Univ., MBA; Diploma in Mgt., AIMA, Dec., 1999]

Hints and Answers

4.32 $\bar{x} = 10.74$ litres per automobile, $\sigma = 5.00$ litres

4.35 (a) Range $= H - L = 18 - 5 = 15$ minutes

Middle 70% of R $= P_{85} - P_{15}$

$$= x_{(85/100) + (1/2)} - x_{(15/100) + (1/2)}$$

$$= x_{(8.5 + 0.5)} - x_{(1.5 + 0.5)} = x_9 - x_2$$

$$= 16 - 5 = 11 \text{ minutes}$$

(b) MAD $= \frac{\sum |x - \bar{x}|}{n} = \frac{47}{10} = 4.7$ minutes

$$s^2 = \frac{\sum x^2 - n\bar{x}^2}{n-1} = \frac{1,431 - 10(10.9)^2}{10-1}$$

$$= 26.99 \text{ minutes}$$

4.36 Lee : $\bar{x} = 62.42$, $s = 0.7497$,

$$CV = (s/\bar{x}) \times 100 = 1.20\%$$

Forest: $\bar{x} = 62.18$, $s = 0.9257$,

$$CV = (s/\bar{x}) \times 100 = 1.49\%$$

Davis : $\bar{x} = 62.46$, $s = 0.9762$,

$$CV = (s/\bar{x}) \times 100 = 1.56\%$$

Based on CV, Lee would be the best customer

4.37 For company A:

$$\bar{x} = A + \frac{\sum fd}{N} \times h = 2200 - \frac{16}{60} \times 200 = 2146.67;$$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times h$$

$$= \sqrt{\frac{88}{60} - \left(\frac{-16}{60}\right)^2} \times 200 = 236.4$$

$$CV = (\sigma/\bar{x}) \times 100 = 11\%$$

For company B: $\bar{x} = 2070$; $\sigma = 158.8$ and $CV = 7.67\%$.

(a) Bulbs of company A are better.

(b) $CV(B) < CV(A)$: Buy company B bulbs as their burning hours are more uniform.

4.38 (a) $\bar{x} = \frac{\sum fm}{n} = \frac{1543}{200} = 7.715$ days;

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{4384.755}{199} = 22.033$$

$$s = \sqrt{22.033} = 4.69 \text{ days}$$

(b) $z = \frac{x - \bar{x}}{s} = \frac{0 - 7.715}{4.69} = -1.644$, i.e. zero day stay is 1.64 standard deviation below the mean, and

$$z = \frac{x - \bar{x}}{s} = \frac{17 - 7.715}{4.69} = 1.97, \text{i.e. 17 days stay in } 1.97 \text{ standard deviation above the mean.}$$

Applying the Chebyshev's theorem with $z = 1.97$, we have

$$\left(1 - \frac{1}{z^2}\right) = \left[1 - \frac{1}{(1.97)^2}\right] = 0.743$$

This indicates that at least 75% patients, i.e. 0.75 (200) = 150 patients should stay between 0 and 17 days.

4.39 (a) $\bar{x} = \frac{\sum x}{n} = 180 \div 5 = 36$ patients

$$s = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{7106 - 5(36)^2}{4}} = \sqrt{156.5} = 12.51 \text{ patients}$$

The middle 75% of data should be in the interval, $\bar{x} \pm 2s = 36 \pm 2(12.51) = (11, 61)$ patients.

(b) $\bar{x} = \frac{\sum x}{n} = 165 \div 5 = 33$ patients

$$s = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{5511 - 5(33)^2}{4}} = \sqrt{16.5} = 4.06 \text{ patients}$$

If distribution is normal, then middle 68% of data should be in the interval $\bar{x} \pm s = 33 \pm 4.06 = (29, 37)$ patients.

Sales Person	\bar{x}	s	$CV = (s/\bar{x}) \times 100$
Raman	88	12.67	14.4%
Sindhu	83.8	6.02	7.2%
Prasad	104.2	16.35	15.7%

Sindhu is most consistent both in terms of s and CV.

4.41 Diameter : 3.97 3.98 3.99 4.00 4.01 4.02 4.03

Frequency: 1 2 4 4 4 4 1

$$\bar{x} = \frac{\sum x}{n} = 80.04 \div 20 = 4.002 \text{ inches}$$

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

$$= \sqrt{\frac{320.325 - 20(4.002)^2}{19}} = 0.016 \text{ inches}$$

If distribution is bell-shaped, then 95% of the gears will have diameters in the interval: $\bar{x} \pm 2s = 4.002 \pm 2(0.016) = (3.970, 4.034)$ inches.

- 4.44** (a) Public : 32; Auto : 32 (b) Public : 4.64; Auto : 1.83 (c) Auto has less variability.

4.45 (i) $\bar{x} = \frac{\sum x}{n}$ or $\sum x = \bar{x}N = 40 \times 100 = 4,000$

Correct $\sum x = 4000 - 50 + 40 = 3990$. Thus
Correct, $\bar{x} = 3990 \div 100 = 39.9$

(ii) $\sigma^2 = \frac{\sum x^2}{N} - (\bar{x})^2$ or $25 = \frac{\sum x^2}{100} - (40)^2$

or $\sum x^2 = 1,62,500$

Correct $\sum x^2 = 1,62,500 - (50)^2 + (40)^2$
 $= 1,62,500 - 2500 + 1600$
 $= 1,61,600$

Correct $\sigma^2 = \frac{\text{Correct } \sum x^2}{N} - (\text{Correct } \bar{x})^2$
 $= \frac{1,61,600}{100} - (39.9)^2 = 23.99$

4.46 (a) $CV(A) = \frac{\sigma}{\bar{x}} \times 100 = \frac{\sqrt{625}}{3200} \times 100 = 0.781$;

$CV(B) = \frac{\sqrt{729}}{2800} \times 100 = 0.964$

There is greater variation in the distribution of wages per employee in factory B.

(b) $\bar{x} = \frac{\sum x}{N}$ or $\sum x = N\bar{x} = 150 \times 2800 = 4,20,000$

Correct $\sum x = 4,20,000 - 3050 + 3650 = 4,20,600$;

Correct, $\bar{x} = \frac{4,20,600}{150} = 2,804$

Variance, $\sigma^2 = \frac{\sum x^2}{N} - (\bar{x})^2$

or $729 = \frac{\sum x^2}{150} - (2800)^2$

or $\sum x^2 = 1,17,61,09,350$

Correct $\sum x^2 = 1,17,61,09,350 - (3050)^2$
 $+ (3650)^2 = 1,18,01,29,350$

Correct $\sigma^2 = \frac{\text{Correct } \sum x^2}{N} - (\text{Correct } \bar{x})^2$
 $= \frac{1,18,01,29,350}{150} - (2804)^2 = 5113$

- 4.47** (a) Total weekly wages: Factory A = $460 \times 100 =$
 $Rs 46,000$; Factory B = $490 \times 80 = Rs 39,200$
 Factory A pays a larger amount.

(b) $CV(\text{Factory A}) = \frac{\sigma}{\bar{x}} \times 100 = \frac{50}{460} \times 100 = 10.87\%$;

$CV(\text{Factory B}) = \frac{40}{490} \times 100 = 8.16\%$

Since $CV(A) > CV(B)$, factory A shows greater variability in wages.

(c) $\bar{x}_{12} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} = \frac{100 \times 460 + 80 \times 490}{100 + 80}$
 $= Rs 473.33$

$$\sigma_{12}^2 = \frac{N_1(\sigma_1^2 + d_1^2) + N_2(\sigma_2^2 + d_2^2)}{N_1 + N_2};$$

$$= \frac{100\{(50)^2 + (13.33)^2\} + 80\{(40)^2 + (16.67)^2\}}{100 + 80}$$

$$= 48.19$$

$d_1 = |\bar{x}_1 - \bar{x}_{12}| = |460 - 473.33| = 13.33$

$d_2 = |\bar{x}_2 - \bar{x}_{12}| = |490 - 473.33| = 16.67$

4.48 $\bar{x} = \frac{\sum x}{N}$ or $\sum x = N\bar{x} = 5 \times 4.4 = 22$

Let two terms x_1 and x_2 are missing. Then
 $x_1 + x_2 + 1 + 2 + 6 = 22$ or $x_1 + x_2 = 13$

Also $\sigma^2 = \frac{\sum x^2}{N} - (\bar{x})^2$ or $8.24 = \frac{\sum x^2}{5} - (4.4)^2$

or $\sum x^2 = 138$

$$\therefore \sum x^2 = x_1^2 + x_2^2 + (1)^2 + (2)^2 + (6)^2$$

$$= 138 \text{ or } x_1^2 + x_2^2 = 97$$

Now $x_1^2 + x_2^2 = (x_1 + x_2)^2 - 2x_1x_2$

or $97 = (13)^2 - 2x_1x_2$ or $x_1x_2 = 36$

$(x_1 - x_2)^2 = x_1^2 + x_2^2 - 2x_1x_2 = 97 - 2(36) = 25$,

or $x_1 - x_2 = 5$

Solving two equations $x_1 + x_2 = 13$ and $x_1 - x_2 = 5$, we have $x_1 = 9$ and $x_2 = 4$.

4.49 QD = $\frac{2}{3}\sigma = \frac{2}{3} \times 5 = \frac{10}{3}$;

$QD = \frac{Q_3 - Q_1}{2} = \frac{10}{3}$ or $Q_3 - Q_1 = \frac{20}{3} = 6.67$

Thus interquartile range is 6.67

4.51 $\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times h$

or $13.1 = \sqrt{\frac{36}{20} - \left(\frac{-6}{20}\right)^2} \times h$ or $h = 10$

$\bar{x} = A + \frac{\sum fd}{N} \times h$

$$\text{or } 107 = A - \frac{6}{20} \times 10 \text{ or } A = 110 \text{ (assumed mean)}$$

Since deviations are taken from $A = 110$ and class interval is, $h = 10$, therefore the class corresponding to $d = 0$ will be 105–115. Other classes will be:

Class :

75–85 85–95 95–105 105–115 115–125 125–135

Frequency :

1	3	4	7	3	2
---	---	---	---	---	---

4.52 Correct $\bar{x} = 39.9$ and $\sigma = 4.9$

$$\text{4.53 } \bar{x} = \frac{\sum x}{N} \text{ or } \sum x = N\bar{x} = 10 \times 9.5 = 95$$

Adding the 11th observation,

We get $\sum x = 95 + 15 = 110$.

$$\text{Then } \bar{x} = \frac{\sum x}{N} = 110 \div 11 = 10$$

$$\text{Also, } \sigma^2 = \frac{\sum x^2}{N} - (\bar{x})^2 \text{ or } (2.5)^2 = \frac{\sum x^2}{10} - (9.5)^2$$

$$\text{or } \sum x^2 = 965$$

Now value of $\sum x^2 = 965 + (15)^2 = 1190$. Then

$$\sigma^2 = \frac{\sum x^2}{N} - (\bar{x})^2 = \frac{1190}{11} - (10)^2 = 2.86$$

4.54 (a) $CV(A) = 30$, $CV(B) = 20$ and $CV(C) = 25$; Share B is more stable.

(b) Dispose share A because of high variability in its price.

4.55 Given $N_1 + N_2 + N_3 = 200$, $N_1 = 50$, $N_3 = 90$, therefore $N_2 = 60$

$$\bar{x}_{123} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2 + N_3\bar{x}_3}{N_1 + N_2 + N_3}$$

$$\text{or } 116 = \frac{50 \times 113 + 60 \times \bar{x}_2 + 90 \times 115}{200}$$

$$\text{or } \bar{x}_2 = 120$$

$$\sigma_{123}^2 = \frac{N_1(\sigma_1^2 + d_1^2) + N_2(\sigma_2^2 + d_2^2) + N_3(\sigma_3^2 + d_3^2)}{N_1 + N_2 + N_3}$$

$$60 = \frac{50\{(6)^2 + (-3)^2\} + 60\{(7)^2 + (4)^2\} + 90\{(\sigma_3^2 + (-1)^2\}}{200}$$

$$\text{or } \sigma_3^2 = 64 \text{ or } \sigma_3 = 8; \text{ where } d_1 = \bar{x}_1 - \bar{x}_{123}; d_2 = \bar{x}_2 - \bar{x}_{123}; d_3 = \bar{x}_3 - \bar{x}_{123}$$

4.56 (a) Firm B pays more wages;

(b) Firm B has greater variability in individual wages

(c) $\bar{x}_{12} = 47.29$ and $\sigma_{12} = 10.605$

Case Studies

Case 4.1: Himgiri Hospital

The hospital recently has installed a new computer-based, interactive, hospital communication system. The system fully integrates the communication activities of admitting, nursing, physician services, laboratory, radiology, pharmacy and assorted medication services, business office, medical records, central supply, dietary services, emergency, and outpatient.

In special training sessions with physicians who were to use the system, the director of the hospital observed that one of the key variables affecting the physicians was the ‘waiting time’ they experienced between inputting data or information requests at a video matrix terminal and the response by the main-frame computer. One of the doctors who is a cardiologist was particularly vocal in his complaints about the system: ‘Look, I can’t wait all day for a machine. I need information that is accurate and in a form I can use. You can’t expect me to also spend time learning how to use your machine—I have enough to do.’

To the physicians, sitting at a terminal and waiting for the computer to respond was simply ‘intoler-

able.’ The director of the hospital was sympathetic to the physicians’ attitude and had negotiated a contract with the computer hardware vendor specifying that the average waiting time not to exceed 10 seconds.

After the system has been operating nearly 15 months, the director conducted a full-scale evaluation. In general, all aspects of the system looked either good or excellent with the exception that only about 60 percent of the physicians were actually using it, and over the past several months there had been a number of complaints about excessive waiting times.

The director was considering the possibility of holding a new series of training sessions for the physicians, but he decided to first review the data collected on actual waiting times experienced by the physicians. These sets of data were available: those collected during the original training session in January 2003 and those collected by staff analysts in March 2004 and May, 2004. These waiting-time (in seconds) data are given below:

<i>January 2003</i>			<i>March 2004</i>				<i>May 2004</i>			
9	8	5	8	6	14	12	7	7	15	16
6	6	7	12	8	7	10	13	7	17	15
9	9	8	12	10			14	7		
7										

Questions for Discussion

1. Calculate the mean waiting time for each of the three sets of data. Do the mean waiting times

appear to be in conformance with the established standard?

2. Calculate the median waiting times for each of the three sets of data. What general conclusions can you draw?
3. Determine the range and standard deviation for each of the three sets of data and consider the implications of the results.

This page is intentionally left blank.

... is touched by that dark
miracle of chance which
makes new magic in a dusty
world.

—Thomas Wolfe

Fundamentals of Probability

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- help yourself to understand the amount of uncertainty that is involved before making important decisions.
- understand fundamentals of probability and various probability rules that help you to measure uncertainty involving uncertainty.
- perform several analyses with respect to business decision involving uncertainty.

5.1 INTRODUCTION

So far we discussed several methods of summarizing sample data to gain knowledge about the entire population or process. Making inferences from sample data, however, involve *uncertainties*. Thus, decision-makers always face some degree of risk while selecting a particular course of action or strategy to solve a decision problem involving uncertainty. It is because each strategy can lead to a number of different possible outcomes (or results). Thus, it is necessary for the decision-makers to enhance their capability of grasping the probabilistic situation so as to gain a deeper understanding of the decision problem and base their decisions on rational considerations. The knowledge of the concepts of probability, probability distributions, and various related statistical techniques is therefore needed. The knowledge of probability and its various types of distributions helps in the development of probabilistic decision models. The material in this chapter is designed to

- (i) explain the fundamentals of probability and related concepts, and
- (ii) illustrate the application of these concepts to decision problems.

5.2 CONCEPTS OF PROBABILITY

In order to obtain a deeper understanding of probability, it is necessary to use certain terms and definitions more precisely. A special type of phenomenon known as *randomness* or *random variation* is of fundamental importance in probability theory. Based upon situations where randomness is present, we can define particular types of occurrences or *events*.

5.2.1 Random Experiment

Random experiment:
A process of obtaining information through observation or measurement of a phenomenon whose outcome is subject to chance.

Random experiment (also called *act*, *trial*, *operation* or *process*) is an activity that leads to the occurrence of one and only one of several possible outcomes which is not likely to be known until its completion, that is, the outcome is not perfectly predictable. This process has the properties that (i) all possible outcomes can be specified in advance, (ii) it can be repeated, and (iii) the same outcome may not occur on various repetitions so that the actual outcome is not known in advance. The variation among experimental outcomes caused by the effects of uncontrolled factors is called *random variation*. It is assumed that these effects vary randomly and unpredictably from one repetition of an experiment to the next.

The outcome (observation or measurement) generated by an experiment may or may not produce a numerical value. Few examples of experiments are as follows:

- (i) Measuring blood pressure of a group of individuals,
- (ii) checking an automobile's petrol mileage,
- (iii) Tossing a coin and observing the face that appears.
- (iv) Testing a product to determine whether it is defective or an acceptable product.
- (v) Measuring daily rainfall, and so on.

In all such cases, there is uncertainty surrounding the outcome until an outcome is observed. For example, if we toss a coin, the outcome will not be known with certainty until either the head or the tail is observed. The number of outcomes may be finite or infinite depending on the nature of the experiment. For example, in the experiment of tossing a coin, the outcomes are finite and are represented by the head and tail, whereas in the experiment of measuring the time between successive failures of an electronic device, the outcomes are infinite and are represented by the time of failure.

The outcome of an experiment may be expressed in numerical or non-numerical value. For example,

- (i) counting the number of arrivals at a service window (numerical outcome), and
- (ii) payment made by cash, cheque, or credit card (non-numerical outcome).

Although an individual outcome associated with a random experiment cannot be predicted exactly, the frequency of occurrence of such an outcome can be noted in a large number of repetitions and thus becomes the basis for resolving problems dealing with uncertainty.

Each experiment may result in one or more outcomes, which are called **events** and denoted by capital letters.

5.2.2 Sample Space

Sample space: The set of all possible outcomes or simple events of an experiment.

The set of all possible distinct outcomes (events) for a random experiment is called the **sample space** (or *event space*) provided.

- (i) no two or more of these outcomes can occur simultaneously;
- (ii) exactly one of the outcomes must occur, whenever the experiment is performed.

Sample space is denoted by the capital letter S.

Illustrations

- Consider the experiment of recording a person's blood type. The four possible outcomes are the following simple events:

$$\begin{array}{ll} E_1 : \text{Blood type A} & E_2 : \text{Blood type B} \\ E_3 : \text{Blood type AB} & E_4 : \text{Blood type O} \end{array}$$

The sample space is $S = \{E_1, E_2, E_3, E_4\}$.

Some experiments can be generated in stages and sample space can be displayed in a *tree diagram*. Each successive level of branching on the tree corresponds to a step required to generate the final outcome as shown in Fig. 5.1. The sample events in the tree diagram form the sample space $S = \{E_1, E_2, E_3, \dots, E_8\}$.

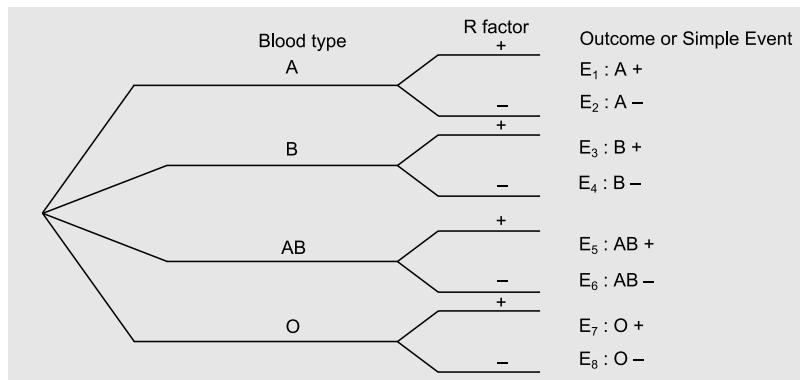


Figure 5.1
Tree Diagram

2. Consider the experiment of tossing two coins. The four possible outcomes are the following sample events

$$E_1 : HH \quad E_2 : HT \quad E_3 : TH \quad E_4 : TT$$

The sample space is $S = \{E_1, E_2, E_3, E_4\}$. The sample events can be displayed in a tree diagram shown in Fig. 5.2.

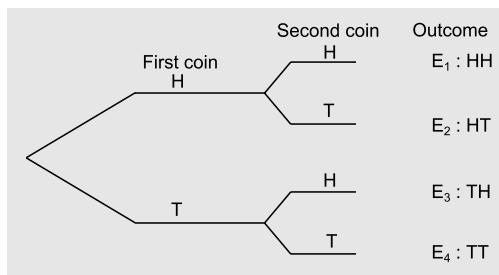


Figure 5.2
Tree Diagram

5.2.3 Event Types

A single possible outcome (or result) of an experiment is called a simple (or elementary) event. An **event** is the set (or collection) of one or more simple events of an experiment in the sample space and having a specific common characteristic. For example, for the above-defined sample space S , the collection $(H, T), (T, H)$ is the event containing simple event as: H or T. Other examples of events are:

- More than 5 customers at a service facility in one hour
- Telephone calls lasting no more than 10 minutes
- 75 per cent marks or better in an examination
- Sales volume of a retail store more than Rs 2,000 on a given day

Mutually Exclusive Events If two or more events cannot occur simultaneously in a single trial of an experiment, then such events are called mutually exclusive events or disjoint events. In other words, two events are mutually exclusive if the occurrence of one of them prevents or rules out the occurrence of the other. For example, the numbers 2 and 3 cannot occur simultaneously on the roll of a dice.

Symbolically, a set of events $\{A_1, A_2, \dots, A_n\}$ is mutually exclusive if $A_i \cap A_j = \emptyset$ ($i \neq j$). This means the intersection of two events is a null set (\emptyset); it is impossible to observe an event that is common in both A_i and A_j .

Collectively Exhaustive Events A list of events is said to be collectively exhaustive when all possible events that can occur from an experiment includes every possible outcome. That is, two or more events are said to be collectively exhaustive if one of the events must occur. Symbolically, a set of events $\{A_1, A_2, \dots, A_n\}$ is collectively exhaustive if the union of these events is identical with the sample space S . That is,

$$S = \{A_1 \cup A_2 \cup \dots \cup A_n\}$$

For example, being a male and female are mutually exclusive and collectively exhaustive events. Similarly, the number 7 cannot come upon the uppermost face during the

Event: Any subset of outcomes of an experiment.

Mutually exclusive events: Events which cannot occur together or simultaneously.

Collectively exhaustive events: The list of events that represents all possible experimental outcomes.

experiment of rolling a dice because the number of faces uppermost has the sample space $S = \{1, 2, 3, 4, 5, 6\}$.

Independent and Dependent Events Two events are said to be *independent* if information about one tells nothing about the occurrence of the other. In other words, outcome of one event does not affect, and is not affected by, the other event. The outcomes of successive tosses of a coin are independent of its preceding toss. Increase in the population (in per cent) per year in India is independent of increase in wheat production (in per cent) per year in the USA.

However, two or more events are said to be dependent if information about one tells something about the other. That is, dependence between characteristics implies that a relationship exists, and therefore, knowledge of one characteristic is useful in assessing the occurrence of the other. For example, drawing of a card (say a queen) from a pack of playing cards without replacement reduces the chances of drawing a queen in the subsequent draws.

Compound Events When two or more events occur in connection with each other, then their simultaneous occurrence is called a compound event. These events may be (i) independent, or (ii) dependent.

Equally Likely Events Two or more events are said to be equally likely if each has an equal chance to occur. That is, one of them cannot be expected to occur in preference to the other. For example, each number may be expected to occur on the uppermost face of a rolling die the same number of times in the long run.

Complementary Events If E is any subset of the sample space, then its complement denoted by (read as E -bar) contains all the elements of the sample space that are not part of E . If S denotes the sample space then

$$\begin{aligned}\bar{E} &= S - E \\ &= \{\text{All sample elements not in } E\}\end{aligned}$$

For example, if E represents companies with sales less than or equal to Rs 25 lakh, written as $E = \{x : x \leq 25\}$, then this set is a complement of the set, $\bar{E} = \{x : x > 25\}$. Obviously such events must be mutually exclusive and collectively exhaustive.

5.3 DEFINITION OF PROBABILITY

Probability: A numerical measure of the likelihood of occurrence of an uncertain event.

A general definition of probability states that **probability** is a numerical measure (between 0 and 1 inclusively) of the likelihood or chance of occurrence of an uncertain event. However, it does not tell us how to compute the probability. In this section, we shall discuss different conceptual approaches of calculating the probability of an event.

5.3.1 Classical Approach

This approach of defining the probability is based on the assumption that all the possible outcomes (finite in number) of an experiment are mutually exclusive and equally likely. It states that, during a random experiment, if there are ' a ' possible outcomes where the favourable event A occurs and ' b ' possible outcomes where the event A does not occur, and all these possible outcomes are mutually exclusive, exhaustive, and equiprobable, then the probability that event A will occur is defined as

$$P(A) = \frac{a}{a+b} = \frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}} = \frac{c(A)}{c(S)}$$

For example, if a fair die is rolled, then on any trial, each event (face or number) is equally likely to occur since there are six equally likely exhaustive events, each will occur $1/6$ of the time, and therefore the probability of any one event occurring is $1/6$. Similarly for the process of selecting a card at random, each event or card is mutually exclusive, exhaustive, and equiprobable. The probability of selecting any one card on a trial is equal to $1/52$, since there are 52 cards. Hence, in general, for a random experiment with

n mutually exclusive, exhaustive, equiprobable events, the probability of any of the events is equal to $1/n$.

Since the probability of occurrence of an event is based on prior knowledge of the process involved, therefore this approach is often called *a priori classical probability approach*. This means, we do not have to perform random experiments to find the probability of occurrence of an event. This also implies that no experimental data are required for computation of probability. Since the assumption of equally likely simple events can rarely be verified with certainty, therefore this approach is not used often other than in games of chance.

The assumption that all possible outcomes are equally likely may lead to a wrong calculation of probability in case some outcomes are more or less frequent in occurrence. For example, if we classify two children in a family according to their sex, then the possible outcomes in terms of number of boys in the family are 0, 1, 2. Thus, according to the **classical approach**, the probability for each of the outcomes should be $1/3$. However, it has been calculated that the probabilities are approximately $1/4$, $1/2$, and $1/4$ for 0, 1, 2 boys respectively. Similarly, we cannot apply this approach to find the probability of a defective unit being produced by a stable manufacturing process as there are only two possible outcomes, defective or non-defective.

Classical approach: The probability of an event A is the ratio of the number of outcomes in favour of A to the number of all possible outcomes, provided experimental outcomes are equally likely to occur.

5.3.2 Relative Frequency Approach

In situations where the outcomes of a random experiment are not all equally likely or when it is not known whether outcomes are equally likely, application of the classical approach is not desirable to quantify the possible occurrence of a random event. For example, it is not possible to state in advance, without repetitive trials of the experiment, the probabilities in cases like (i) whether a number greater than 3 will appear when die is rolled or (ii) if a lot of 100 items will include 10 defective items.

Relative frequency approach: The probability of an event A is the ratio of the number of times that A has occurred in n trials of an experiment.

This approach of computing probability is based on the assumption that a random experiment can be repeated a large number of times under identical conditions where trials are independent to each other. While conducting a random experiment, we may or may not observe the desired event. But as the experiment is repeated many times, that event may occur some proportion of time. Thus, the approach calculates *the proportion of the time (i.e. the relative frequency) with which the event occurs over an infinite number of repetitions of the experiment under identical conditions*. Since no experiment can be repeated an infinite number of times, therefore a probability can never be exactly determined. However, we can approximate the probability of an event by recording the relative frequency with which the event has occurred over a finite number of repetitions of the experiment under identical conditions. For example, if a die is tossed n times and s denotes the number of times the event A (i.e., number 4, 5, or 6) occurs, then the ratio $P(A) = c(s)/n$ gives the proportion of times the event A occurs in n trials, and are also called relative frequencies of the event in n trials. Although our estimate about $P(A)$ may change after every trial, yet we will find that the proportion $c(s)/n$ tends to cluster around a unique central value as the number of trials n becomes even larger. This unique central value (also called probability of event A) is defined as:

$$P(A) = \lim_{n \rightarrow \infty} \left\{ \frac{c(s)}{n} \right\}$$

where $c(s)$ represents the number of times that an event s occurs in n trials of an experiment.

Since the probability of an event is determined objectively by repetitive empirical observations of experimental outcomes, it is also known as *empirical probability*. Few situations to which this approach can be applied are follows:

- (i) Buying lottery tickets regularly and observing how often you win
- (ii) Commuting to work daily and observing whether or not a certain traffic signal is red when you cross it.

- (iii) Observing births and noting how often the baby is a female
- (iv) Surveying many adults and determining what proportion smokes.

Subjective approach:

The probability of an event based on the personal beliefs of an individual.

5.3.3 Subjective Approach

The **subjective approach** of calculating probability is always based on the degree of beliefs, convictions, and experience concerning the likelihood of occurrence of a random event. It is thus a way to quantify an individual's beliefs, assessment, and judgment about a random phenomenon. Probability assigned for the occurrence of an event may be based on just guess or on having some idea about the relative frequency of past occurrences of the event. This approach must be used when either sufficient data are not available or sources of information giving different results are not known.

5.3.4 Fundamental Rules of Probability

No matter which approach is used to define probability, the following fundamental rules must be satisfied. Let S be the sample space of an experiment that is partitioned into mutually exclusive and exhaustive events A_1, A_2, \dots, A_n which may be elementary or compound. The probability of any event A in S is governed by the following rules:

- (i) Each probability should fall between 0 and 1, i.e. $0 \leq P(A_i) \leq 1$, for all i , where $P(A_i)$ is read as: 'probability of event A_i '. In other words, the probability of an event is restricted to the range *zero to one* inclusive, where zero represents an impossible event and one represents a certain event.

For example, probability of the number seven occurring, on rolling a dice, $P(7) = 0$, because this number is an impossible event for this experiment.

- (ii) $P(S) = P(A_1) + P(A_2) + \dots + P(A_n) = 1$, where $P(S)$ is read as: 'probability of the certain event'. This rule states that the sum of probabilities of all simple events constituting the sample space is equal to one. This also implies that if a random experiment is conducted, one of its outcomes in its sample space is certain to occur.

Similarly, the probability of an impossible event or an empty set is zero. That is $P(\emptyset) = 0$.

- (iii) If events A_1 and A_2 are two elements in S and if occurrence of A_1 implies that A_2 occurs, that is, if A_1 is a subset of A_2 , then the probability of A_1 is less than or equal to the probability of A_2 . That is, $P(A_1) \leq P(A_2)$.

- (iv) $P(\bar{A}) = 1 - P(A)$, that is, the probability of an event that does not occur is equal to one minus the probability of the event that does occur (the probability rule for complementary events).

5.3.5 Glossary of Probability Terms

If A and B are two events, then

$A \cup B$ = an event which represents the occurrence of either A or B or both.

$A \cap B$ = an event which represents the simultaneous occurrence of A and B .

\bar{A} = complement of event A and represents non-occurrence of A .

$\bar{A} \cap \bar{B}$ = both A and B do not occur.

$\bar{A} \cap B$ = event A does not occur but event B occurs.

$A \cap \bar{B}$ = event A occurs but event B does not occur.

$(A \cap \bar{B}) \cup (\bar{A} \cap B)$ = exactly one of the two events A and B occurs.

5.4 COUNTING RULES FOR DETERMINING THE NUMBER OF OUTCOMES

In order to assign probabilities to experimental outcomes, it is first necessary to identify and then count them. Following are three important rules for counting the experimental outcomes.

5.4.1 Multistep Experiments

The counting rule for multistep experiments helps us to determine the number of experimental outcomes without listing them. The rule is defined as:

If an experiment is performed in k stages with n_1 ways to accomplish the first stage, n_2 ways to accomplish the second stage ... and n_k ways to accomplish the k th stage, then the number of ways to accomplish the experiment is $n_1 \times n_2 \times \dots \times n_k$.

Illustrations

- Tossing of two coins can be thought of as a two-step experiment in which each coin can land in one of two ways: head (H) and tail (T). Since the experiment involves two steps, forming the pair of faces (H or T), the total number of simple events in S will be

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

The elements of S indicate that there are $2 \times 2 = 4$ possible outcomes.

When the number of alternative events in each of the several trials is same, that is, $n_1 = n_2 = \dots = n_k$, then the multi step method gives $n_1 \times n_2 \times \dots \times n_k = n^k$.

For example, if the coins involved in a coin-tossing experiment are four, then the number of experimental outcomes will be $2 \times 2 \times 2 \times 2 = 2^4 = 16$.

- Suppose a person can take three routes from city A to city B, four from city B to city C and three from city C to city D. Then the possible routes for reaching from city A to D, while he must travel from A to B to C to D are : (A to B) \times (B to C) \times (C to D) = $3 \times 4 \times 3 = 36$ ways.

5.4.2 Combinations

Sometimes the ordering or arrangement of objects is not important, but only the objects that are chosen. For example, (i) you may not care in what order the books are placed on the shelf, but only which books you are able to shelve. (ii) When a five-person committee is chosen from a group of 10 students, the order of choice is not important because all 5 students will be equal members of committee.

This counting rule for combinations allows us to select r (say) number of outcomes from a collection of n distinct outcomes without caring in what order they are arranged. This rule is denoted by

$$C(n, r) = {}^nC_r = \frac{n!}{r!(n-r)!}$$

where $n! = n(n - 1)(n - 2) \dots 3 \cdot 2 \cdot 1$ and $0! = 1$.

The notation ! means factorial, for example, $4! = 4 \times 3 \times 2 \times 1 = 24$.

Important Results

- ${}^nC_r = {}^nC_{n-r}$ and ${}^nC_n = 1$.
- If n objects consist of all n_1 of one type, all n_2 of another type, and so on upto n_k of the k th type, then the total number of selections that can be made of 1, 2, 3 upto n objects is $(n_1 + 1)(n_2 + 1) \dots (n_k + 1) - 1$.
- The total number of selections from n objects all different is $2^n - 1$.

5.4.3 Permutations

This rule of counting involves ordering or permutations. This rule helps us to compute the number of ways in which n distinct objects can be arranged, taking r of them at a time.

The total number of permutations of n objects taken r at a time is given by

$$P(n, r) = {}^nP_r = \frac{n!}{(n-r)!}$$

By permuting each combination of r objects among themselves, we shall obtain all possible permutations of n objects, r at a time. Each combination gives rise to $r!$ permutations, so that $r! C(n, r) = P(n, r) = n!/(n-r)!$.

Example 5.1: Of ten electric bulbs, three are defective but it is not known which are defective. In how many ways can three bulbs be selected? How many of these selections will include at least one defective bulb?

Solution: Three bulbs out of 10 bulbs can be selected in ${}^{10}C_3 = 120$ ways. The number of selections which include exactly one defective bulb will be ${}^7C_2 \times {}^3C_1 = 63$.

Similarly, the number of selections which include exactly two and three defective bulbs will be ${}^7C_1 \times {}^3C_2 = 21$ and ${}^3C_3 = 1$, respectively. Thus, the total number of selections including at least one defective bulb is $63 + 21 + 1 = 85$.

Example 5.2: A bag contains 6 red and 8 green balls.

- If one ball is drawn at random, then what is the probability of the ball being green?
- If two balls are drawn at random, then what is the probability that one is red and the other green?

Solution: (a) Since the bag contains 6 red and 8 green balls, therefore it contains $6 + 8 = 14$ equally likely outcomes, that is, $S = \{r, g\}$. But one ball out of 14 balls can be drawn in ways, that is,

$${}^{14}C_1 = \frac{14!}{1!(14-1)!} = 14 \text{ ways}$$

Let A be the event of drawing a green ball. Then, out of these 8 green balls, one green ball can be drawn in 8C_1 ways:

$${}^8C_1 = \frac{8!}{1!(8-1)!} = 8$$

Hence, $P(A) = \frac{c(A)}{c(S)} = \frac{8}{14}$

$$(b) \text{ All exhaustive number of cases, } c(S) = {}^{14}C_2 = \frac{14!}{2!(14-2)!} = 91.$$

Also, out of 6 red balls, one red ball can be drawn in 6C_1 ways and out of 8 green balls, one green ball can be drawn in 8C_1 ways. Thus, the total number of favourable cases is:

$$c(B) = {}^6C_1 \times {}^8C_1 = 6 \times 8 = 48$$

Thus, $P(B) = \frac{c(B)}{c(S)} = \frac{48}{91}$

Example 5.3: Tickets are numbered from 1 to 100. They are well shuffled and a ticket is drawn at random. What is the probability that the drawn ticket has

- an even number?
- the number 5 or a multiple of 5?
- a number which is greater than 75?
- a number which is a square?

Solution: Since any of the 100 tickets can be drawn, therefore exhaustive number of cases are $c(S) = 100$.

(a) Let A be the event of getting an even numbered tickets. Then, $c(A) = 50$, and hence

$$P(A) = 50/100 = 1/2$$

(b) Let B be the event of getting a ticket bearing the number 5 or a multiple of 5, that is,

$$B = [5, 10, 15, 20, \dots, 95, 100]$$

which are 20 in number, $c(B) = 20$. Thus, $P(B) = 20/100 = 1/5$.

(c) Let C be the event of getting a ticket bearing a number greater than 75, that is,

$$C = \{76, 77, \dots, 100\}$$

which are 25 in number, $c(C) = 25$. Thus, $P(C) = 25/100 = 1/4$.

- (d) Let D be the event of getting a ticket bearing a number which is a square, that is,

$$D = \{1, 4, 9, 16, 25, 36, 49, 64, 81, 100\}$$

which are 10 in number, $c(D) = 10$. Thus, $P(D) = 10/100 = 1/10$.

Conceptual Questions 5A

1. (a) Discuss the different schools of thought on the interpretation of probability. How does each school define probability?
[HP Univ., MBA, 1998; Delhi Univ., MBA, 1999]
1. (b) Describe briefly the various schools of thought on probability. Discuss its importance in business decision-making.
[Delhi Univ., MBA, 1999; Kumaon Univ., 2000]
2. Explain what you understand by the term probability. Discuss its importance in business decision-making.
[Delhi Univ., MBA, 2002]
3. (a) Give the classical and statistical definitions of probability and state the relationship, if any, between the two definitions.
(b) Critically examine the '*a priori*' definition of probability showing clearly the improvement which the empirical version of probability makes over it.
4. Define independent and mutually exclusive events. Can two events be mutually exclusive and independent simultaneously? Support your answer with an example.
[Sukhadia Univ., MBA; Delhi Univ., MBA, 1999]
5. Compare and contrast the three interpretations of probability.
6. Explain the difference between statistically independent and statistically dependent events.
7. Explain the meaning of each of the following terms:

(a) Random phenomenon	(b) Statistical experiment
(c) Random event	(d) Sample space
8. What do you mean by probability? Explain the importance of probability. [Madras Univ., MA(Eco), MBA, 2003]
9. State the multiplicative theorem of probability. How is the result modified when the events are independent.
10. Life insurance premiums are higher for older people, but auto insurance premiums are generally higher for younger people. What does this suggest about the risks and probabilities associated with these two areas of insurance business?
11. Distinguish between the two concepts in each of the following pairs:
 - (a) Elementary event and compound events
 - (b) Mutually exclusive events and overlapping events
 - (c) Sample space and sample point
12. (a) Define the terms—joint probability, marginal probability, and conditional probability
(b) By comparing the three kinds of probabilities (joint, conditional, and marginal), explain what information is provided by each
13. Suppose an entire shipment of 1000 items is inspected and 50 items are found to be defective. Assume the defective items are not removed from the shipment before being sent to a retail outlet for sale. If you purchase one item from this shipment, what is the probability that it will be one of the defective items?
14. Suppose you are told that the price of a particular stock will increase with a probability of 0.7.
 - (a) How is this probability interpreted?
 - (b) Assuming the definition of probability in terms of long-run relative frequencies, how would you find the probability that a stock price will increase?

Self-Practice Problems 5A

- 5.1 Three unbiased coins are tossed. What is the probability of obtaining:

(a) all heads	(b) two heads
(c) one head	(d) at least one head
(e) at least two heads,	(f) all tails
- 5.2 A card is drawn from a well-shuffled deck of 52 cards. Find the probability of drawing a card which is neither a heart nor a king.
- 5.3 In a single throw of two dice, find the probability of getting (a) a total of 11, (b) a total of 8 or 11, and (c) same number on both the dice.
- 5.4 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 they are all graduates? What is the probability of at least one graduate?
- 5.5 A bag contains 25 balls numbered 1 through 25. Suppose an odd number is considered a 'success'. Two balls are drawn from the bag with replacement. Find the probability of getting

(a) two successes	(b) exactly one success	(c) at least one success	(d) no successes
-------------------	-------------------------	--------------------------	------------------
- 5.6 A bag contains 5 white and 8 red balls. Two drawings of 3 balls are made such that (a) the balls are replaced before the second trial, and (b) the balls are not replaced before the second trial. Find the probability that the first drawing will give 3 white and the second, 3 red balls in each case. [MD Univ., BCom; GND Univ., MA 1995; Kerala Univ., MCom, 1998]

- 5.7** Three groups of workers contain 3 men and one woman, 2 men and 2 women, and 1 man and 3 women respectively. One worker is selected at random from each group. What is the probability that the group selected consists of 1 man and 2 women?

[Nagpur Univ., MCom, 1997]

- 5.8** What is the probability that a leap year, selected at random, will contain 53 Sundays?

[Agra Univ., MCom; Kurukshetra Univ., MCom, 1996;
MD Univ., MCom, 1998]

- 5.9** A university has to select an examiner from a list of 50 persons, 20 of them women and 30 men, 10 of them knowing Hindi and 40 not, 15 of them being teachers and the remaining 35 not. What is the probability of the university selecting a Hindi-knowing woman teacher?

[Jammu Univ., MCom, 1997]

Hints and Answers

- 5.1** (a) $P(\text{all heads}) = 1/8$ (b) $P(\text{two heads}) = 3/8$
 (c) $P(\text{one head}) = 3/8$ (d) $P(\text{at least one head}) = 7/8$
 (e) $P(\text{at least two heads}) = 4/8 = 1/2$
 (f) $P(\text{all tails}) = 1/8$.

$$\text{5.2 } P(\text{neither a heart nor a king}) = \frac{36C_1}{52C_1} = \frac{36}{52}$$

$$\text{5.3 } c(S) = 36; \quad P(\text{total of 11}) = 2/36 \quad P(\text{total of 9 or 11}) \\ = 7/36$$

$$\text{5.4 } P(\text{all graduate}) = \frac{5C_3 \times 15C_0}{20C_3} = \frac{10 \times 1}{1140} = \frac{1}{114}$$

$$P(\text{no graduate}) = \frac{15C_3 \times 5C_0}{20C_2} = \frac{455 \times 1}{1140} = \frac{91}{28}$$

$$P(\text{at least one graduate}) = 1 - \frac{91}{228} = \frac{137}{228}$$

$$\text{5.5 } (a) \quad P(\text{two successes}) = \frac{13}{25} \times \frac{13}{25} = \frac{169}{625}$$

$$(b) \quad P(\text{exactly one success}) = \frac{13}{25} \times \frac{12}{25} + \frac{13}{25} \times \frac{12}{25} = \frac{312}{625}$$

$$(c) \quad P(\text{at least one success}) = P(\text{exactly one success}) \\ + P(\text{two successes}) = \frac{312}{625} + \frac{169}{625} = \frac{481}{625}$$

$$(d) \quad P(\text{no successes}) = \frac{12}{25} \times \frac{12}{25} = \frac{144}{625}$$

- 5.6** (a) *When balls are replaced:*

Total number of balls in the bag = $5 + 8 = 13$.

3 balls can be drawn from 13 in ${}^{13}C_3$ ways;

3 white balls can be drawn from 5 in 5C_3 ways;

3 red balls can be drawn from 8 in 8C_3 ways.

The probability of 3 red balls in the second trial

$$= \frac{5C_3}{13C_3} = \frac{5}{143}$$

Probability of 3 red balls in the second trial

$$= \frac{4C_2}{12C_2} = \frac{28}{143}$$

The probability of the compound event

$$\frac{5}{143} \times \frac{28}{143} = \frac{140}{20449} = 0.007$$

- (b) *When balls are not replaced:*

At the first trial, 3 white balls can be drawn in 5C_3 ways.

The probability of drawing three white balls at the

$$\text{first trial} = \frac{5C_3}{13C_3} = \frac{5}{143}$$

When the white balls have been drawn and not replaced, the bag contains 2 white and 8 red balls. Therefore, at the second trial, 3 balls can be drawn from 10 in ${}^{10}C_3$ ways and 3 red balls can be drawn from 8 in 8C_3 ways.

The probability of 3 red balls in the second trial

$$= \frac{8C_3}{10C_3} = \frac{7}{15}$$

The probability of the compound event

$$= \frac{5}{142} \times \frac{7}{15} = \frac{7}{429} = 0.016.$$

- 5.7** There are three possibilities in this case:

- (i) Man is selected from the 1st group and women from 2nd and 3rd groups; or
- (ii) Man is selected from the 2nd group and women from the 1st and 3rd groups; or
- (iii) Man is selected from the 3rd group and women from 1st and 2nd groups.

The probability of selecting a group of 1 man and 2 women is:

$$\left(\frac{3}{4} \times \frac{2}{4} \times \frac{3}{4} \right) + \left(\frac{2}{4} \times \frac{1}{4} \times \frac{3}{4} \right) + \left(\frac{1}{4} \times \frac{1}{4} \times \frac{2}{4} \right) \\ = \frac{9}{32} + \frac{3}{32} + \frac{1}{32} = \frac{13}{32}.$$

- 5.8** A leap year consists of 366 days, therefore it contains 52 complete weeks and 2 extra days. These 2 days may make the following 7 combinations:

- (i) Monday and Tuesday
- (ii) Tuesday and Wednesday
- (iii) Wednesday and Thursday
- (iv) Thursday and Friday
- (v) Friday and Saturday

- (vi) Saturday and Sunday
 (vii) Sunday and Monday

Of these seven equally likely cases, only the last two are favourable. Hence the required probability is $2/7$.

5.9 Probability of selecting a woman = $20/50$;

Probability of selecting a teacher = $15/50$
 Probability of selecting a Hindi-knowing candidate = $10/50$
 Since the events are independent, the probability of the university selecting a Hindi-knowing woman teacher = $(20/50) \times (15/50) \times (10/50) = 3/125$.

5.5 RULES OF PROBABILITY AND ALGEBRA OF EVENTS

In probability, we use set theory notations to simplify the presentation of ideas. As discussed earlier in this chapter, the probability of the occurrence of an event A is expressed as:

$$P(A) = \text{probability of event } A \text{ occurrence}$$

Such single probabilities are called **marginal (or unconditional) probabilities** because it is the probability of a single event occurring. In the coin tossing example, the marginal probability of a tail or head in a toss can be stated as $P(T)$ or $P(H)$.

Marginal probability: The unconditional probability of an event occurring.

5.5.1 Rules of Addition

The addition rules are helpful when we have two events and are interested in knowing the probability that at least one of the events occurs.

Mutually Exclusive Events The rule of addition for mutually exclusive (disjoint), exhaustive, and equally likely events states that

If two events A and B are mutually exclusive, exhaustive, and equiprobable, then the probability of either event A or B or both occurring is equal to the sum of their individual probabilities.

This rule is expressed in the following formula

$$\begin{aligned} P(\text{A or B}) &= P(A \cup B) = \frac{n(A \cup B)}{n(S)} = \frac{n(A) + n(B)}{n(S)} \\ &= \frac{n(A)}{n(S)} + \frac{n(B)}{n(S)} = P(A) + P(B) \end{aligned} \quad (5-1)$$

where $A \cup B$ (read as 'A union B') denotes the union of two events A and B and it is the set of all sample points belonging to A or B or both. This rule can also be illustrated by the **Venn diagram** shown in Fig. 5.3. Here two circles contain all the sample points in events A and B. The overlap of the circles indicates that some sample points are contained in both A and B.

Illustration: Consider the pattern of arrival of customers at a service counter during the first hour it is open along with its probability:

No. of persons :	0	1	2	3	4 or more
Probability :	0.1	0.2	0.3	0.3	0.1

To understand the probability that either 2 or 3 persons will be there during the first hour, we have

$$P(2 \text{ or } 3) = P(2) + P(3) = 0.3 + 0.3 = 0.6$$

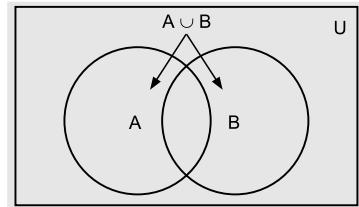
The formula (5-1) can be expanded to include more than two events. In particular, if there are n events in a sample space that are mutually exclusive, then the probability of the union of these events is given by

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n) \quad (5-2)$$

For example, if we are interested in knowing the probability that there will be two or more persons during the first hour, then using formula (5-2), we have

$$\begin{aligned} P(2 \text{ or more}) &= P(2, 3, 4 \text{ or more}) = P(2) + P(3) + P(4) \\ &= 0.3 + 0.3 + 0.1 = 0.7 \end{aligned}$$

Figure 5.3
 Union of Two Events



Venn diagram: A pictorial representation for showing the sample space and operations involving events. The sample space is represented by a rectangle and events as circles.

An important special case of formula (5-1) is for complementary events. Let A be any event and \bar{A} be the complement of A. Obviously A and \bar{A} are mutually exclusive and exhaustive events. Thus, either A occurs or it does not, is given by

$$\begin{aligned} P(A \text{ or } \bar{A}) &= P(A) + P(\bar{A}) = P(A) + \{1 - P(A)\} = 1 \\ \text{or} \quad P(A) &= 1 - P(\bar{A}) \end{aligned} \quad (5-3)$$

For example, if a dice is rolled, then the probability whether an odd number of spots occurs or does not.

Partially Overlapping (or Joint) Events If events A and B are not mutually exclusive, it is possible for both events to occur simultaneously? This means these events have some sample points in common. Such events are also called *joint* (or *overlapping*) events. The sample points in common (belong to both events) represent the joint event $A \cap B$ (read as: A intersection B). The addition rule in this case is stated as:

If two events A and B are not mutually exclusive, then the probability of either A or B or both occurring is equal to the sum of their individual probabilities minus the probability of A and B occurring together.

This rule is expressed in the following formula:

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) - P(A \text{ and } B) \\ \text{or} \quad P(A \cup B) &= P(A) + P(B) - P(A \cap B) \end{aligned} \quad (5-4)$$

This addition rule can also be illustrated by the Venn-diagram shown in Fig. 5.4.

Figure 5.4
Partially Overlapping Events

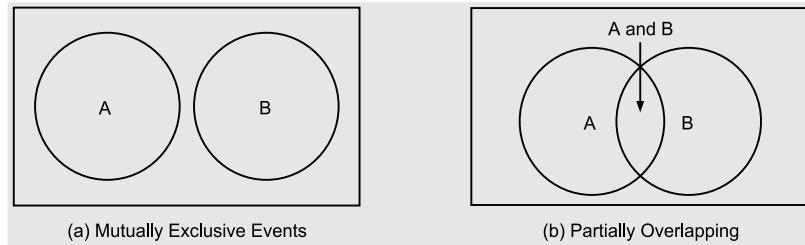


Illustration: Suppose 70 per cent of all tourists who come to India will visit Agra while 60 per cent will visit Goa and 50 per cent of them will visit both Agra and Goa. The probability that a tourist will visit either Goa or Agra or both is obtained by applying formula (5-4) as follows:

$$\begin{aligned} P(\text{Agra or Goa}) &= P(\text{Agra}) + P(\text{Goa}) - P(\text{both Agra and Goa}) \\ &= 0.70 + 0.60 - 0.50 = 0.8 \end{aligned}$$

Consequently, the probability that a tourist will visit neither Agra nor Goa is calculated by

$$P(\text{neither Agra nor Goa}) = 1 - P(\text{Agra or Goa}) = 1 - 0.80 = 0.20$$

The formula (5-4) can be expanded to include more than two events. In particular, if there are three events that are not mutually exclusive, then the probability of the union of these events is given by

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) \\ &\quad - P(C \cap A) + P(A \cap B \cap C) \end{aligned} \quad (5-5)$$

Remark: The rules of addition are applicable for calculating probability of events in case of simultaneous trials.

Example 5.4: What is the probability that a randomly chosen card from a deck of cards will be either a king or a heart.

Solution: Let event A and B be the king and heart in a deck of 52 cards, respectively. Then, it is given that

Card	Probability	Reason
King	$P(A) = 4/52$	4 kings in a deck of 52 cards
Heart	$P(B) = 13/52$	13 hearts in a deck of 52 cards
King of heart	$P(A \text{ and } B) = 1/52$	1 King of heart in a deck of 52 cards

Using the formula (5-4), we get

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) - P(A \text{ and } B) \\ &= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = 0.3077 \end{aligned}$$

Example 5.5: Of 1000 assembled components, 10 have a working defect and 20 have a structural defect. There is a good reason to assume that no component has both defects. What is the probability that randomly chosen component will have either type of defect?

[Delhi Univ., MBA, 2003]

Solution: Let the event A and B be the component which has working defect and has structural defect, respectively. Then it is given that

$$P(A) = 10/1000 = 0.01, P(B) = 20/1000 = 0.02 \text{ and } P(A \text{ and } B) = 0$$

The probability that a randomly chosen component will have either type of defect is given by

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) - P(A \text{ and } B) \\ &= 0.01 + 0.02 - 0.0 = 0.03 \end{aligned}$$

Example 5.6: A survey of 200 retail grocery shops revealed following monthly income pattern:

Monthly Income (Rs)	Number of Shops
Under Rs 20,000	102
20,000 to 30,000	61
30,000 and above	37

- (a) What is the probability that a particular shop has monthly income under Rs 20,000
- (b) What is the probability that a shop selected at random has either an income between Rs 20,000 and Rs 30,000 or an income of Rs. 30,000 and more?

Solution: Let the events A, B and C represent the income under three categories, respectively.

(a) Probability that a particular shop has monthly income under Rs 20,000 is $P(A) = 102/200 = 0.51$.

(b) Probability that shop selected at random has income between Rs 20,000 and Rs 30,000 or Rs 30,000 and more is given by

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) \\ &= \frac{61}{200} + \frac{37}{200} = 0.305 + 0.185 = 0.49 \end{aligned}$$

Example 5.7: From a sales force of 150 persons, one will be selected to attend a special sales meeting. If 52 of them are unmarried, 72 are college graduates, and $3/4$ of the 52 that are unmarried are college graduates, find the probability that the sales person selected at random will be neither single nor a college graduate.

Solution: Let A and B be the events that a sales person selected is married and that he is a college graduate, respectively. Then, it is given that

$$P(A) = 52/150, P(B) = 72/150; P(A \text{ and } B) = (3/4)(52/150) = 39/150$$

The probability that a salesperson selected at random will be neither single nor a college graduate is:

$$\begin{aligned} P(\bar{A} \cap \bar{B}) &= 1 - P(A \cup B) = 1 - \{P(A) + P(B) - P(A \cap B)\} \\ &= 1 - \left\{ \frac{52}{150} + \frac{72}{150} - \frac{39}{150} \right\} = \frac{13}{30} \end{aligned}$$

Example 5.8: From a computer tally based on employer records, the personnel manager of a large manufacturing firm finds that 15 per cent of the firm's employees are supervisors and 25 per cent of the firm's employees are college graduates. He also discovers that 5 per cent are both supervisors and college graduates. Suppose an employee is selected at random from the firm's personnel records, what is the probability of:

- (a) selecting a person who is both a college graduate and a supervisor?
- (b) selecting a person who is neither a supervisor nor a college graduate?

Solution: Let A and B be the events that the person selected is a supervisor and that he is a college graduate, respectively. Given that

$$P(A) = 15/100; P(B) = 25/100; P(A \text{ and } B) = 5/100$$

- (a) Probability of selecting a person who is both a college graduate and a supervisor is:
 $P(A \text{ and } B) = 5/100 = 0.05$

- (b) Probability of selecting a person who is neither a supervisor nor a college graduate is:

$$\begin{aligned} P(\bar{A} \text{ and } \bar{B}) &= 1 - P(A \cup B) = 1 - [P(A) + P(B) - P(A \cap B)] \\ &= 1 - \left(\frac{15}{100} + \frac{25}{100} - \frac{5}{100} \right) = \frac{65}{100} = 0.65 \end{aligned}$$

Example 5.9: The probability that a contractor will get a plumbing contract is $2/3$ and the probability that he will not get an electrical contract is $5/9$. If the probability of getting at least one contract is $4/5$, what is the probability that he will get both?

Solution: Let A and B denote the events that the contractor will get a plumbing and electrical contract, respectively. Given that

$$P(A) = 2/3; P(B) = 1 - (5/9) = 4/9; P(A \cup B) = 4/5$$

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) = \frac{2}{3} + \frac{4}{9} - \frac{4}{5} = \frac{14}{45} = 0.31$$

Thus, the probability that the contractor will get both the contracts is 0.31.

Example 5.10: An MBA applies for a job in two firms X and Y. The probability of his being selected in firm X is 0.7 and being rejected at Y is 0.5. The probability of at least one of his applications being rejected is 0.6. What is the probability that he will be selected by one of the firms?

Solution: Let A and B denote the event that an MBA will be selected in firm X and will be rejected in firm Y, respectively. Then, given that

$$P(A) = 0.7, P(\bar{A}) = 1 - 0.7 = 0.3$$

$$P(B) = 0.5, P(\bar{B}) = 1 - 0.5 = 0.5, P(\bar{A} \cup \bar{B}) = 0.6$$

$$\text{Since } P(A \cap B) = 1 - P(\bar{A} \cup \bar{B}) = 1 - 0.6 = 0.4$$

therefore, probability that he will be selected by one of the firms is given by

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.7 + 0.5 - 0.4 = 0.8 \end{aligned}$$

Thus, the probability of an MBA being selected by one of the firms is 0.8.

5.5.2 Rules of Multiplication

Statistically Independent Events When the occurrence of an event does not affect and is not affected by the probability of occurrence of any other event, the event is said to be a *statistically independent event*. There are three types of probabilities under statistical independence: *marginal*, *joint*, and *conditional*.

- **Marginal Probability:** A marginal or unconditional probability is the simple probability of the occurrence of an event. For example, in a fair coin toss, the outcome of each toss is an event that is statistically independent of the outcomes of every other toss of the coin.
- **Joint Probability:** The probability of two or more independent events occurring together or in succession is called the **joint probability**. The joint probability of two or more independent events is equal to the product of their marginal probabilities. In particular, if A and B are independent events, the probability that both A and B will occur is given by

$$P(AB) = P(A \cap B) = P(A) \times P(B) \quad (5-6)$$

Suppose we toss a coin twice. The probability that in both the cases the coin will turn up head is given by

$$P(H_1 H_2) = P(H_1) \times P(H_2) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

The formula (5-6) is applied here because the probability of any outcome is not affected by any preceding outcome, in other words, outcomes are independent.

- **Conditional Probability:** It is the probability of a particular event occurring, given that another event has occurred. The **conditional probability** of event A, given that event B has already occurred is written as: $P(A|B)$. Similarly, we may write $P(B|A)$. The vertical bar is read as ‘given’ and events appearing to the right of the bar are those that you know have occurred. Two events A and B are said to be independent if and only $P(A|B) = P(A)$ or $P(B|A) = P(B)$. Otherwise, events are said to be dependent.

Joint probability: The probability of two events occurring together or in succession.

Conditional probability: The probability of an event occurring, given that another event has occurred.

Since, in the case of independent events the probability of occurrence of either of the events does not depend or affect the occurrence of the other, therefore in the coin tossing example, the probability of a head occurrence in the second toss, given that head resulted in the first toss, is still 0.5. That is, $P(H_2 | H_1) = 0.5 = P(H_2)$. It is because of the fact that the probabilities of heads and tails are the same for every toss and in no way influenced by whether it was a head or tail which occurred in the previous toss.

Statistically Dependent Events When the probability of an event is dependent upon or affected by the occurrence of any other event, the events are said to be **statistically dependent**. There are three types of probabilities under statistical dependence: *joint*, *conditional*, and *marginal*.

- **Joint Probability:** If A and B are dependent events, then the joint probability as discussed under statistical dependence case is no longer equal to the product of their respective probabilities. That is, for dependent events

$$P(A \text{ and } B) = P(A \cap B) \neq P(A) \times P(B)$$

Accordingly, $P(A) \neq P(A | B)$ and $P(B) \neq P(B | A)$

The joint probability of events A and B occurring together or in succession under statistical dependence is given by

$$P(A \cap B) = P(A) \times P(B | A)$$

or $P(A \cap B) = P(B) \times P(A | B)$

- **Conditional Probability:** Under statistical dependence, the conditional probability of event B, given that event A has already occurred, is given by

$$P(B | A) = \frac{P(A \cap B)}{P(A)}$$

Similarly, the conditional probability of A, given that event B has occurred, is

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

Statistical dependence: The condition when the probability of occurrence of an event is dependent upon, or affected by, the occurrence of some other event.

- **Marginal Probability:** The marginal probability of an event under statistical dependence is the same as the marginal probability of an event under statistical independence.

The marginal probability of events A and B can be written as:

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap \bar{B}) \\ \text{and} \quad P(B) &= P(A \cap B) + P(\bar{A} \cap B) \end{aligned}$$

Example 5.11: The odds against student X solving a Business Statistics problem are 8 to 6, and odds in favour of student Y solving the problem are 14 to 16.

- (a) What is the chance that the problem will be solved if they both try independently of each other?
- (b) What is the probability that none of them is able to solve the problem?

[Delhi Univ., MBA, 1998]

Solution: Let A = The event that the first student solves the problem,

B = The event that the second student solves the problem.

$$P(A) = \frac{6}{8+6} = \frac{6}{14} \quad \text{and} \quad P(B) = \frac{14}{14+16} = \frac{14}{30}$$

- (a) Probability that the problem will be solved

$$\begin{aligned} &= P(\text{at least one of them solves the problem}) \\ &= P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \\ &= P(A) + P(B) - P(A) \times P(B) [\text{because the events are independent}] \\ &= \frac{6}{14} + \frac{14}{30} - \frac{6}{14} \times \frac{14}{30} = \frac{73}{105} = 0.695 \end{aligned}$$
- (b) Probability that neither A nor B solves the problem

$$\begin{aligned} P(\bar{A} \text{ and } \bar{B}) &= P(\bar{A}) \times P(\bar{B}) \\ &= [1 - P(A)] \times [1 - P(B)] = \frac{8}{14} \times \frac{16}{30} = \frac{32}{105} = 0.305 \end{aligned}$$

Example 5.12: The probability that a new marketing approach will be successful is 0.6. The probability that the expenditure for developing the approach can be kept within the original budget is 0.50. The probability that both of these objectives will be achieved is 0.30. What is the probability that at least one of these objectives will be achieved. For the two events described above, determine whether the events are independent or dependent.

[Delhi Univ., MBA, 2003]

Solution: Let A = The event that the new marketing approach will be successful
 B = The event that the expenditure for developing the approach can be kept within the original budget

Given that $P(A) = 0.60$, $P(B) = 0.50$ and $P(A \cap B) = 0.30$

Probability that both events A and B will be achieved is given by

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.60 + 0.50 - 0.30 = 0.80 \end{aligned}$$

If events A and B are independent, then their joint probability is given by

$$P(A \cap B) = P(A) \times P(B) = 0.60 \times 0.50 = 0.30$$

Since this value is same as given in the problem, events are independent.

Example 5.13: A piece of equipment will function only when the three components A, B, and C are working. The probability of A failing during one year is 0.15, that of B failing is 0.05, and that of C failing is 0.10. What is the probability that the equipment will fail before the end of the year?

Solution: Given that

$$\begin{aligned} P(\text{A failing}) &= 0.15; P(\text{A not failing}) = 1 - P(\text{A}) = 0.85 \\ P(\text{B failing}) &= 0.05; P(\text{B not failing}) = 1 - P(\text{B}) = 0.95 \\ P(\text{C failing}) &= 0.10; P(\text{C not failing}) = 1 - P(\text{C}) = 0.90 \end{aligned}$$

Since all the three events are independent, therefore the probability that the equipment will work is given by

$$P(\bar{A} \cap \bar{B} \cap \bar{C}) = P(\bar{A}) \times P(\bar{B}) \times P(\bar{C}) = 0.85 \times 0.95 \times 0.90 = 0.726$$

Probability that the equipment will fail before the end of the year is given by

$$\begin{aligned} P(A \cup B \cup C) &= 1 - P(\bar{A} \cap \bar{B} \cap \bar{C}) \\ &= 1 - P(\bar{A}) \cdot P(\bar{B}) \cdot P(\bar{C}) \\ &= 1 - \{0.85 \times 0.95 \times 0.90\} = 1 - 0.726 = 0.274 \end{aligned}$$

Example 5.14: A market research firm is interested in surveying certain attitudes in a small community. There are 125 households broken down according to income, ownership of a telephone, and ownership of a TV.

	Households with Annual Income of Rs 8000 or Less		Households with Annual Income Above Rs 8000		Total
	Telephone subscriber	No telephone	Telephone subscriber	No telephone	
	Own TV set	27	20	18	10
No TV set	18	10	12	10	50
Total	45	30	30	20	125

- (a) What is the probability of getting a TV owner in a random draw?
- (b) If a household has an income of over Rs 8000 and is a telephone subscriber, what is the probability that he owns a TV?
- (c) What is the conditional probability of drawing a household that owns a TV, given that the household is a telephone subscriber?
- (d) Are the events ‘ownership of a TV’ and ‘telephone subscriber’ statistically independent? Comment. [Himachal Univ., MBA, 1998]

Solution: (a) Probability of drawing a TV owner at random,

$$P(\text{TV owner}) = 75/125 = 0.6$$

(b) There are $30(18 + 12)$ persons whose household income is above Rs 8000 and are also telephone subscribers. Out of these, 18 own TV sets. Hence, the probability of this group of persons having a TV set, is: $18/30 = 0.6$.

(c) Out of $75(27 + 18 + 18 + 12)$ households who are telephone subscribers, $45(27 + 18)$ households have TV sets. Hence, the conditional probability of drawing a household that owns a TV given that the household is a telephone subscriber is: $45/75 = 0.6$.

(d) Let A and B be the events representing TV owners and telephone subscribers respectively. The probability of a person owning a TV, $P(A) = 75/125$. The probability of a person being a telephone subscriber, $P(B) = 75/125$.

The probability of a person being a telephone subscriber as well as a TV owner is:

$$P(A \text{ and } B) = 45/125 = 9/25$$

But $P(A) \times P(B) = (75/125)(75/125) = 9/25$

Since $P(AB) = P(A) \times P(B)$, therefore, we conclude that the events ‘ownership of a TV’ and ‘telephone subscriber’ are statistically independent.

Example 5.15: A company has two plants to manufacture scooters. Plant I manufactures 80 per cent of the scooters and Plant II manufactures 20 per cent. In plant I, only 85 out of 100 scooters are considered to be of standard quality. In plant II, only 65 out of 100 scooters are considered to be of standard quality. What is the probability that a scooter selected at random came from plant I, if it is known that it is of standard quality? [Madras Univ., MCom, 1996; Delhi Univ., MBA, 1998]

Solution: Let A = The scooter purchased is of standard quality

B = The scooter is of standard quality and came from plant I

C = The scooter is of standard quality and came from plant II

D = The scooter came from plant I

The percentage of scooters manufactured in plant I that are of standard quality is 85 per cent of 80 per cent, that is, $0.85 \times (80/100) = 68$ per cent or $P(B) = 0.68$.

The percentage of scooters manufactured in plant II that are of standard quality is 65 per cent of 20 per cent, that is, $0.65 \times (20 \div 100) = 13$ per cent or $P(C) = 0.13$.

The probability that a customer obtains a standard quality scooter from the company is, $0.68 + 0.13 = 0.81$.

The probability that the scooters selected at random came from plant I, if it is known that it is of standard quality, is given by

$$P(D|A) = \frac{P(D \text{ and } A)}{P(A)} = \frac{0.68}{0.81} = 0.84$$

Example 5.16: A husband and wife appear in an interview for two vacancies in the same post. The probability of husband's selection is $1/7$ and that of wife's selection is $1/5$. What is the probability that

- (a) both of them will be selected,
- (b) only one of them will be selected, and
- (c) none of them will be selected.

[Bharthidasan Univ., MCom, 1996; Delhi Univ., MBA, 1999]

Solution: Let A and B be the events of the husband's and wife's selection, respectively. Given that $P(A) = 1/7$ and $P(B) = 1/5$.

(a) The probability that both of them will be selected is:

$$P(A \text{ and } B) = P(A) P(B) = (1/7) \times (1/5) = 1/35 = 0.029$$

(b) The probability that only one of them will be selected is:

$$\begin{aligned} P[(A \text{ and } \bar{B}) \text{ or } (B \text{ and } \bar{A})] &= P(A \text{ and } \bar{B}) + P(B \text{ and } \bar{A}) \\ &= P(A) P(\bar{B}) + P(B) P(\bar{A}) \\ &= P(A) [1 - P(B)] + P(B) [1 - P(A)] \\ &= \frac{1}{7} \left(1 - \frac{1}{5}\right) + \frac{1}{5} \left(1 - \frac{1}{7}\right) = \left(\frac{1}{7} \times \frac{4}{5}\right) + \left(\frac{1}{5} \times \frac{6}{7}\right) \\ &= \frac{10}{35} = 0.286 \end{aligned}$$

(c) The probability that none of them will be selected is:

$$P(\bar{A}) \times P(\bar{B}) = (6/7) \times (4/5) = 24/35 = 0.686$$

Example 5.17: The odds that A speaks the truth is 3 : 2 and the odds that B speaks the truth is 5 : 3. In what percentage of cases are they likely to contradict each other on an identical point? [Delhi Univ., MBA, 1999]

Solution: Let X and Y denote the events that A and B speak truth, respectively. Given that

$$P(X) = 3/5; \quad P(\bar{X}) = 2/5; \quad P(Y) = 5/8; \quad P(\bar{Y}) = 3/8$$

The probability that A speaks the truth and B speaks a lie is: $(3/5)(3/8) = 9/40$

The probability that B speaks the truth and A speaks a lie is: $(5/8)(2/5) = 10/40$

So the compound probability is: $\frac{9}{40} + \frac{10}{40} = \frac{19}{40}$

Hence, percentage of cases in which they contradict each other is $(19/40) \times 100 = 47.5$ per cent.

Example 5.18: The data for the promotion and academic qualification of a company is given below:

Promotional Status	Academic Qualification		Total
	MBA(A)	Non-MBA	
Promoted (B)	0.14	0.26	0.40
Non-promoted	0.21	0.39	0.60
Total	0.35	0.65	1.00

- (a) Calculate the conditional probability of promotion after an MBA has been identified.
 (b) Calculate the conditional probability that it is an MBA when a promoted employee has been chosen.
 (c) Find the probability that a promoted employee was an MBA. [IGNOU, 1995]

Solution: It is given that, $P(A)=0.35$, $P=0.65$, $P(B)=0.40$, $P=0.60$, and $P(A \cap B)=0.14$

- (a) The probability of being ‘promoted’ after an MBA employee has been identified is:

$$P(B | A) = \frac{P(A \cap B)}{P(A)} = \frac{0.14}{0.35} = 0.34$$

(b) If a promoted employee has been chosen, then the probability that the person is an MBA is:

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{0.14}{0.40} = 0.35$$

- (c) The probability that a promoted employee was an MBA is:

$$P(A \cap B) = P(A) \times P(B | A) = 0.35 \times 0.34 = 0.12$$

or $P(A \cap B) = P(B) \times P(A | B) = 0.40 \times 0.35 = 0.12$

Example 5.19: The probability that a trainee will remain with a company is 0.6. The probability that an employee earns more than Rs 10,000 per month is 0.5. The probability that an employee who is a trainee remained with the company or who earns more than Rs 10,000 per month is 0.7. What is the probability that an employee earns more than Rs 10,000 per month given that he is a trainee who stayed with the company?

Solution: Let A and B be the events that a trainee who remained with the company and the event that an employee earns more than Rs 10,000, respectively. Given that

$$P(A) = 0.6, P(B) = 0.5, \text{ and } P(A \cup B) = P(A \cup B) = 0.7$$

The probability that an employee earns more than Rs 10,000, given that he is trainee who remained with the company, is given by

$$P(B | A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(A \cap B)}{P(A)}$$

We know that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$,

or $P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.6 + 0.5 - 0.7 = 0.4$

Hence the required probability is:

$$P(B | A) = \frac{(A \cap B)}{P(A)} = \frac{0.4}{0.6} = 0.667$$

Example 5.20: Two computers A and B are to be marketed. A salesman who is assigned the job of finding customers for them has 60 per cent and 40 per cent chances of succeeding for computers A and B, respectively. The two computers can be sold independently. Given that he was able to sell at least one computer, what is the probability that computer A has been sold? [IGNOU, MBA, 2002; Delhi Univ., MBA, 1999, 2002]

Solution: Let us define the events

$$E_1 = \text{Computer A is marketed and } E_2 = \text{Computer B is marketed.}$$

It is given that $P(E_1) = 0.60$, $P(E_2) = 0.40$

$$P(E_1 \text{ and } E_2) \text{ or } P(E_1 \cap E_2) = P(E_1) \cdot P(E_2) = 0.60 \times 0.40 = 0.24$$

Hence, the probability that computer A has been sold given that the salesman was able to sell at least one computer is given by

$$\begin{aligned} P(E_1 | E_1 \cup E_2) &= \frac{P\{E_1 \cap (E_1 \cup E_2)\}}{P(E_1 \cup E_2)} = \frac{P(E_1)}{P(E_1 \cup E_2)} \\ &= \frac{P(E_1)}{P(E_1) + P(E_2) - P(E_1 \cap E_2)} = \frac{0.60}{0.60 + 0.40 - 0.24} \\ &= \frac{0.60}{0.76} = 0.789 \end{aligned}$$

Example 5.21: A study of job satisfaction was conducted for four occupations: Cabin maker lawyer, doctor and systems analyst. Job satisfaction was measured on a scale of 0–100. The data obtained are summarized in the following table:

Occupation	Under 50	50–59	60–69	70–79	80–89	Total
Cabin maker	0	2	4	3	1	10
Lawyer	6	2	1	1	0	10
Doctor	0	5	2	1	2	10
Systems Analyst	2	1	4	3	0	10
	8	10	11	8	3	40

- (a) Develop a joint probability table.
- (b) What is the probability of one of the participants studied had a satisfaction score in 80's?
- (c) What is the probability of a satisfaction score in the 80's, given the study participant was a doctor?
- (d) What is the probability of one of the participants studied was a lawyer.
- (e) What is the probability of one of the participants was a lawyer and received a score under 50?
- (f) What is the probability of a satisfaction score under 50 given a person is a lawyer.
- (g) What is the probability of a satisfaction score of 70 or higher?

[Delhi Univ., MBA, 2003]

Solution: (a) Joint probability table is given below

Occupation	Under 50	50–59	60–69	70–79	80–89
Cabin maker	0.000	0.050	0.100	0.075	0.250
Lawyer	0.150	0.050	0.025	0.025	0.250
Doctor	0.000	0.125	0.050	0.025	0.250
System Analyst	0.050	0.025	0.100	0.075	0.250

- (b) $P(\text{Satisfaction score in the 80's}) = 3/40$
- (c) $P(\text{Satisfaction score in 80's, given participant was doctor}) = \frac{2/40}{10/40} = \frac{1}{5}$
- (d) $P(\text{Participant was doctor}) = 10/40$
- (e) $P(\text{Lawyer and score under 50}) = \frac{P(\text{Lawyer} \cap \text{Score under 50})}{P(\text{Score under 50})} = \frac{6}{40}$
- (f) $P(\text{Score under 50 Lawyer}) = \frac{P(\text{Score under 50} \cap \text{Lawyer})}{P(\text{Lawyer})} = \frac{6/40}{10/40} = \frac{6}{10}$
- (g)
$$\begin{aligned} P(\text{Satisfaction score of 70 or higher}) &= P(\text{Score of 70 and above}) + P(\text{Score of 80 and above}) \\ &= \frac{8}{40} + \frac{3}{40} = \frac{11}{40} \end{aligned}$$

Example 5.22: A market survey was conducted in four cities to find out the preference for brand A soap. The responses are shown below:

	Delhi	Kolkata	Chennai	Mumbai
Yes	45	55	60	50
No	35	45	35	45
No opinion	5	5	5	5

- (a) What is the probability that a consumer selected at random, preferred brand A?
- (b) What is the probability that a consumer preferred brand A and was from Chennai?

- (c) What is the probability that a consumer preferred brand A, given that he was from Chennai?
 (d) Given that a consumer preferred brand A, what is the probability that he was from Mumbai? [Delhi Univ., MBA, 2002; Kumaon Univ., MBA, 1999]

Solution: The information from responses during market survey is as follows:

	Delhi	Kolkata	Chennai	Mumbai	Total
Yes	45	55	60	50	210
No	35	45	35	45	160
No opinion	5	5	5	5	20
Total	85	105	100	100	390

Let X denote the event that a consumer selected at random preferred brand A. Then

- (a) The probability that a consumer selected at random preferred brand A is:

$$P(X) = 210/390 = 0.5398$$

- (b) The probability that a consumer preferred brand A and was from Chennai (C) is:

$$P(X \cap C) = 60/390 = 0.1538$$

- (c) The probability that a consumer preferred brand A, given that he was from Chennai:

$$P(X|C) = \frac{P(A \cap C)}{P(C)} = \frac{60/390}{100/390} = \frac{0.153}{0.256} = 0.597$$

- (d) The probability that the consumer belongs to Mumbai, given that he preferred brand A

$$P(M|X) = \frac{P(M \cap X)}{P(X)} = \frac{50/390}{210/390} = \frac{0.128}{0.538} = 0.237$$

Example 5.23: The personnel department of a company has records which show the following analysis of its 200 engineers.

Age	Bachelor's Degree Only	Master's Degree	Total
Under 30	90	10	100
30 to 40	20	30	50
Over 40	40	10	50
Total	150	50	200

If one engineer is selected at random from the company, find:

- (a) The probability that he has only a bachelor's degree.
 (b) The probability that he has a master's degree, given that he is over 40.
 (c) The probability that he is under 30, given that he has only a bachelor's degree.

[Kumaon Univ., MBA 1998]

Solution: Let A, B, C, and D denote the events that an engineer is under 30 years of age, 40 years of age, has bachelor's degree only, and has a master's degree, respectively. Therefore:

- (a) The probability of an engineer who has only a bachelor's degree is:

$$P(C) = 150/200 = 0.75$$

- (b) The probability of an engineer who has a master's degree, given that he is over 40 years is:

$$P(D|B) = \frac{P(D \cap B)}{P(B)} = \frac{10/200}{50/200} = \frac{10}{50} = 0.20$$

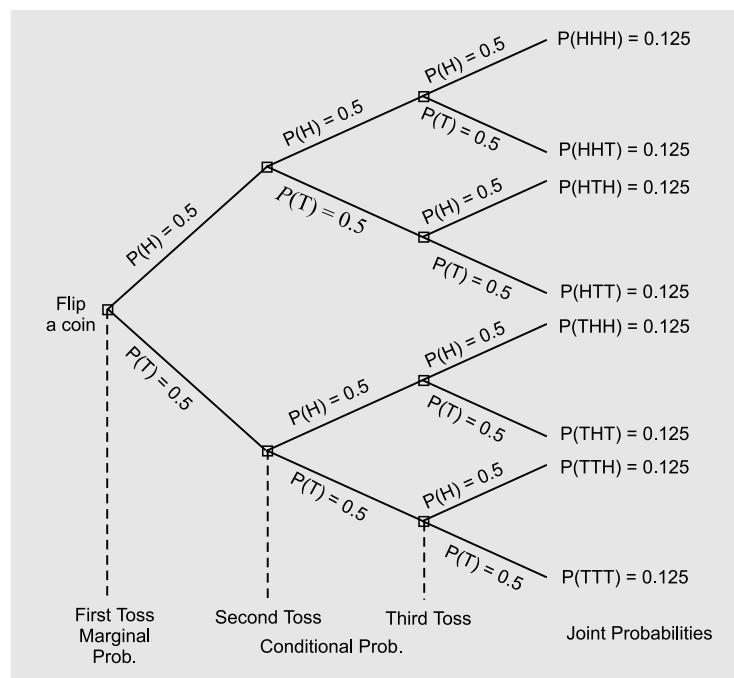
- (c) The probability of an engineer who is under 30 years, given that he has only bachelor's degree is:

$$P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{90/200}{150/200} = \frac{90}{150} = 0.60$$

5.6 PROBABILITY TREE DIAGRAM

Decision-makers at times face difficulty in constructing the joint probability table. Thus, they prefer to use the *probability tree* to probability calculations. The probability tree diagram for tossing a coin three consecutive times is shown in Fig. 5.5.

Figure 5.5
Probability Tree Diagram



The probability tree diagram is convenient for calculating joint probabilities when events occur at different times or stages. In probability trees, time moves from left to right. The complete tree exhibits each outcome as a *single path* from beginning to end. Each path corresponds to a distinct joint event. A joint event is represented by a path through the tree and its probability is determined by multiplying all the individual branch probabilities for its path.

In this example of three tosses of a coin, at each toss the probability of either event's (H or T) occurring remains the same, that is, the events are independent. The joint probabilities of events occurring in succession are calculated by multiplying the probabilities of each event.

The events emanating from a single breaking point are mutually exclusive and collectively exhaustive, so that exactly one must occur. All the probabilities on branches within the same fork must therefore sum to 1.

In this case, the results in the diagram should not be confused with conditional probabilities. The probability of a head and then two tails occurring on three consecutive tosses is computed prior to any tosses taking place. If the first two tosses have already occurred, then the probability of getting a tail on the third toss is still 0.5, that is, $P(T|HT) = P(T) = 0.5$.

Example 5.24: Each salesperson is rated either below average, average, or above average with respect to sales ability. Each of them is also rated with respect to his or her potential for advancement — fair, good or excellent. These traits of the 500 sales person are given below:

Sales Ability	Potential for Advancement		
	Fair	Good	Excellent
Below average	16	12	22
Average	45	60	45
Above average	93	72	135

- (a) What is the probability that a randomly selected salesperson will have above average sales ability and excellent potential for advancement?
- (b) Construct a tree diagram showing all the probabilities — conditional probabilities and joint probabilities.

Solution: (a) Let A and B represent the event that a salesperson will have above average sales ability and excellent potential for advancement. The joint probability of these traits is given by

$$P(A \text{ and } B) = P(A) \times P(B) = \frac{300}{500} \times \frac{135}{300} = 0.27$$

- (b) The tree diagram of probabilities is shown in Fig. 5.6.

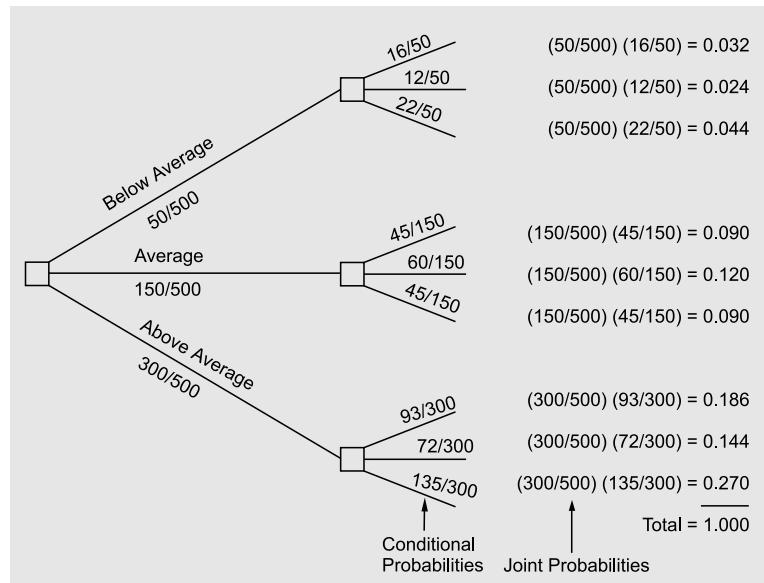


Figure 5.6
Tree Diagram

Self-Practice Problems 5B

- 5.10** Mr. X has 2 shares in a lottery in which there are 2 prizes and 5 blanks. Mr. Y has 1 share in a lottery in which there is 1 prize and 2 blanks. Show that the chance of Mr. X's success to that of Mr. Y's is 15 : 7.
- 5.11** Explain whether or not each of the following claims could be correct:
- A businessman claims that the probability that he will get contract A is 0.15 and that he will get contract B is 0.20. Furthermore, he claims that the probability of getting A or B is 0.50.
 - A market analyst claims that the probability of selling ten million rupees of plastic A or five million rupees of plastic B is 0.60. He also claims that the probability of selling ten million rupees of A and five million rupees of B is 0.45.
- 5.12** The probability that an applicant for a Management Accountant's job has a postgraduate degree is 0.3, he has had some work experience as a chief Financial Accountant is 0.7, and that he has both is 0.2. Out of 300 applicants, approximately, what number would have either a postgraduate degree or some professional work experience?
- 5.13** A can hit a target 3 times in 5 shots; B, 2 times in 5 shots; C, 3 times in 4 shots. They fire a volley. What is the probability that 2 shots hit?
- 5.14** A problem in business statistics is given to five students, A, B, C, D, and E. Their chances of solving it are $1/2$, $1/3$, $1/4$, $1/5$, and $1/6$ respectively. What is the probability that the problem will be solved?
[Madras Univ., BCom, 1996; Kumaon Univ., MBA, 2000]
- 5.15** A husband and wife appear in an interview for two vacancies for the same post. The probability of husband's selection is $1/7$ and that of wife's selection is $1/5$. What is the probability that
- only one of them will be selected?
 - both of them will be selected?
 - none of them will be selected?
- 5.16** A candidate is selected for interviews for three posts. For the first, there are 3 candidates, for the second, there are 4, and for the third, there are 2. What is the probability of his getting selected for at least one post?
- 5.17** Three persons A, B, and C are being considered for appointment as Vice-Chancellor of a university, and whose chances of being selected for the post are in the proportion 14 : 2 : 3 respectively. The probability that A if selected, will introduce democratization in the university structure is 0.3, and the corresponding probabilities for B and C doing the same are respectively 0.5 and 0.8. What is the probability that democratization would be introduced in the university?

- 5.18** There are three brands, say X, Y, and Z of an item available in the market. A consumer chooses exactly one of them for his use. He never buys two or more brands simultaneously. The probabilities that he buys brands X, Y, and Z are 0.20, 0.16, and 0.45, respectively.
- What is the probability that he does not buy any of the brands?
 - Given that a customer buys some brand, what is the probability that he buys brand X?
- 5.19** There is 50-50 chance that a contractor's firm, A, will bid for the construction of a multi-storeyed building. Another firm, B, submits a bid and the probability is 3/5 that it will get the job, provided that firm A does not submit a bid. If firm A submits a bid, the probability that firm B will get the job is only 2/3. What is the probability that firm B will get the job?
- 5.20** Plant I of XYZ manufacturing organization employs 5 production and 3 maintenance foremen, plant II of same organization employs 4 production and 5 maintenance foremen. From any one of these plants, a single selection of two foremen is made. Find the probability that one of them would be a production and the other a maintenance foreman. [Bombay Univ., MMS, 1997]
- 5.21** Two sets of candidates are competing for positions on the board of directors of a company. The probability that the first and second sets will win are 0.6 and 0.4 respectively. If the first set wins, the probability of introducing a new product is 0.8 and the corresponding probability if the second set wins is 0.3.
- What is the probability that the new product will be introduced?
 - If the new product was introduced, what is the probability that the first set won as directors?
- 5.22** If a machine is correctly set up, it will produce 90 per cent acceptable items. If it is incorrectly setup, it will produce 40 per cent acceptable items. Past experience shows that 80 per cent of the setups are correctly done. If after a certain setup, the machine produces 2 acceptable items as the first 2 pieces, find the probability that the machine is correctly set up. [Delhi Univ., BCom, (Hons), 1998]
- 5.23** A firm plans to bid Rs 300 per tonne for a contract to supply 1,000 tonnes of a metal. It has two competitors A and B. It assumes the probability of A bidding less than Rs 300 per tonne to be 0.3 and B's bid to be less than Rs 300 per tonne to be 0.7. If the lowest bidder gets all the business and the firms bid independently, what is the expected value of the contract to the firm?
- 5.24** An investment consultant predicts that the odds against the price of a certain stock going up during the next week are 2 : 1 and odds in favour of the price remaining the same are 1 : 3. What is the probability that the price of the stock will go down during the next week?
- 5.25** An article manufactured by a company consists of two parts A and B. In the process of manufacture of part A, 9 out of 100 are likely to be defective. Similarly, 5 out of 100 are likely to be defective in the manufacture of part B. Calculate the probability that the assembled part will not be defective.
- 5.26** A product is assembled from three components X, Y, and Z, the probability of these components being defective is 0.01, 0.02, and 0.05, respectively. What is the probability that the assembled product will not be defective?
- 5.27** The daily production of a machine producing a very complicated item gives the following probabilities for the number of items produced: $P(1) = 0.20$, $P(2) = 0.35$, and $P(3) = 0.45$. Furthermore, the probability of defective items being produced is 0.02. Defective items are assumed to occur independently. Determine the probability of no defectives during a day's production.
- 5.28** The personnel department of a company has records which show the following analysis of its 200 engineers:
- | Age (Years) | Bachelor's Degree only | Master's Degree | Total |
|-------------|------------------------|-----------------|-------|
| Under 30 | 90 | 10 | 100 |
| 30 to 40 | 20 | 30 | 50 |
| Over 40 | 40 | 10 | 50 |
| | 150 | 50 | 200 |
- If one engineer is selected at random from the company, find:
- the probability that he has only a bachelor's degree;
 - the probability that he has a master's degree given that he is over 40;
 - the probability that he is under 30 given that he has only a bachelor's degree.
- [HP Univ., MBA; Kumaon Univ., MBA 1998]
- 5.29** In a certain town, males and females form 50 per cent of the population. It is known that 20 per cent of the males and 5 per cent of the females are unemployed. A research student studying the employment situation selects unemployed persons at random. What is the probability that the person selected is (a) male, (b) female?
- [Delhi Univ. MCom, 1999; Kumaon Univ., MBA, 1998]
- 5.30** You note that your officer is happy in 60 per cent cases of your calls. You have also noticed that if he is happy, he accedes to your requests with a probability of 0.4, whereas if he is not happy, he accedes to your requests with a probability of 0.1. You call on him one day and he accedes to your request. What is the probability of his being happy? [HP, MBA, 1996]
- 5.31** In a telephone survey of 1000 adults, respondents were asked about the expenses on a management education and the relative necessity of some form of financial assistance. The respondents were classified according to whether they currently had a child studying in a school of management and whether they thought that the loan burden for most management students is: too high, right amount, or too little. The proportions responding in each category are given below.
- | | Too High
(A) | Right Amount
(B) | Too Little
(C) |
|------------------------------------|-----------------|---------------------|-------------------|
| Child studying management (D) : | 0.35 | 0.08 | 0.01 |
| No child studying management (E) : | 0.25 | 0.20 | 0.11 |

Suppose one respondent is chosen at random from this group. Then

- What is the probability that the respondent has a child studying management.
 - Given that the respondent has a child studying management, what is the probability that he/she ranks the loan burden as 'too high'.
 - Are events D and A independent? Explain.
- 5.32** In a colour preference experiment, eight toys are placed in a container. The toys are identical except for colour — two are red, and six are green. A child is asked to choose two toys at random. What is the probability that the child chooses the two red toys?
- 5.33** A survey of executives dealt with their loyalty to the company. One of the questions was, 'If you were given an offer by another company equal to or slightly better

than your present position, would you remain with the company?' The responses of 200 executives in the survey cross-classified with their length of service with the company are shown below:

Loyalty	Length of Service					Total
	Less than 1 year	1–5 years	6–10 years	More than 10 years		
Would remain :	10	30	5	75	120	
Would not remain :	25	15	10	30	80	

What is the probability of randomly selecting an executive who is loyal to the company (would remain) and who has more than 10 years of service.

Hints and Answers

- 5.10** Considering Mr. X's chances of success.

A = event that 1 share brings a prize and 1 share goes blank.

B = event that both the shares bring prizes.

C = event that X succeeds in getting atleast one prize
= $A \cup B$.

Since A and B are mutually exclusive, therefore

$$P(C) = P(A \cup B) = P(A) + P(B) = \frac{2C_1 \times 5C_1}{7C_2} + \frac{2C_2 \times 5C_0}{7C_2}$$

Similarly, if D denotes the event that Y succeeds in getting a prize, then we have

$$P(D) = \frac{1C_1}{3C_1} = \frac{1}{3}$$

$$\text{X's chance of success: Y's chance of success} = \frac{15}{21} : \frac{1}{3} = 15 : 7.$$

- 5.11** (a) $P(A \cap B) = -0.15$

(b) $P(A) + P(B) = 1.05$

- 5.12** Let A = Applicant has P.G degree; B = Applicant has work experience;

Given, $P(A) = 0.3$, $P(B) = 0.7$, and $P(A \cap B) = 0.2$. Therefore

$$300 \times P(A \cup B) = 300[P(A) + P(B) - P(A \cap B)] = 240$$

- 5.13** The required event that two shots may hit the target, can happen in the following mutually exclusive cases:

- A and B hit and C fails to hit the target
- A and C hit and B fails to hit the target
- B and C hit and A fails to hit the target

Hence, the required probability that any two shots hit is given by, $P = P(i) + P(ii) + P(iii)$.

Let E_1 , E_2 , and E_3 be the event of hitting the target by A, B, and C respectively. Therefore

$$P(i) = P(E_1 \cap E_2 \cap \bar{E}_3) = P(E_1) \cdot P(E_2) \cdot P(\bar{E}_3) \\ = \left(\frac{3}{5}\right) \cdot \left(\frac{2}{5}\right) \cdot \left(1 - \frac{3}{4}\right) = \frac{6}{100}$$

$$P(ii) = P(E_1 \cap \bar{E}_2 \cap E_3) = \left(\frac{3}{5}\right) \left(1 - \frac{2}{5}\right) \left(\frac{3}{4}\right) = \frac{27}{100}$$

$$P(iii) = P(\bar{E}_1 \cap E_2 \cap E_3) = \left(1 - \frac{3}{5}\right) \left(\frac{2}{5}\right) \left(\frac{3}{4}\right) = \frac{12}{100}$$

Since all the three events are mutually exclusive events, hence the required probability is given by

$$P(i) + P(ii) + P(iii) = \frac{6}{100} + \frac{27}{100} + \frac{12}{100} = \frac{9}{20}$$

- 5.14** P(problem will be solved)

$$= 1 - P(\text{problem is not solved})$$

$$= 1 - P(\text{all students fail to solve the problem})$$

$$= 1 - P(\bar{A} \cap \bar{B} \cap \bar{C} \cap \bar{D} \cap \bar{E})$$

$$= 1 - P(\bar{A}) P(\bar{B}) P(\bar{C}) P(\bar{D}) P(\bar{E})$$

$$= 1 - \left(1 - \frac{1}{2}\right) \left(1 - \frac{1}{3}\right) \left(1 - \frac{1}{4}\right) \left(1 - \frac{1}{5}\right) \left(1 - \frac{1}{6}\right)$$

$$= 1 - \frac{1}{6} = \frac{5}{6}$$

- 5.15** P(only one of them will be selected)

$$= P(H \cap \bar{W}) \cup (\bar{H} \cap W) = P(H \cap \bar{W}) + (\bar{H} \cap W)$$

$$= P(H) P(\bar{W}) + P(\bar{H}) P(W)$$

$$= \frac{1}{7} \left(1 - \frac{1}{5}\right) + \left(1 - \frac{1}{7}\right) \frac{1}{5} = \frac{2}{7}$$

- (b) P(both of them will be selected)

$$P(H \cap W) = P(H) \cdot P(W) = \frac{1}{35}$$

- (c) P(None of them will be selected)

$$P(\bar{H} \cap \bar{W}) = P(\bar{H}) \cdot P(\bar{W}) = \frac{24}{35}$$

- 5.17** P(D) = Prob. of the event that democratization would be introduced

$$= P[(A \cap D) \cup [(B \cap D) \cup (C \cap D)]]$$

$$= P[(A \cap D) + P(B \cap D) + P(C \cap D)]$$

$$\begin{aligned}
 &= P(A) \cdot P(D | A) + P(B) \cdot P(D | B) \\
 &\quad + P(C) \cdot P(D | C) \\
 &= 0.3 \left(\frac{4}{9} \right) + 0.5 \left(\frac{2}{9} \right) + 0.8 \left(\frac{3}{9} \right) = 0.51
 \end{aligned}$$

5.18 (a) $P(\text{customer does not buy any brand})$

$$\begin{aligned}
 &= P[\bar{X} \cap \bar{Y} \cap \bar{Z}] = 1 - P[X \cup Y \cup Z] \\
 &= 1 - [P(X) + P(Y) + P(Z)] \\
 &= 1 - [0.20 + 0.16 + 0.45] = 0.19
 \end{aligned}$$

(b) $P(\text{customer buys brand X}) = P[X | (X \cup Y \cup Z)]$

$$\begin{aligned}
 &= \frac{P[X \cap (X \cup Y \cup Z)]}{P(X \cup Y \cup Z)} \\
 &= \frac{P(X)}{P(X) + P(Y) + P(Z) - P(X \cap Y) - P(Y \cap Z)} \\
 &\quad - P(X \cap Z) + P(X \cap Y \cap Z) \\
 &= \frac{0.2}{0.2 + 0.16 + 0.45 - 0 - 0 - 0} = 0.247
 \end{aligned}$$

5.19 $P(A) = 1/2$, $P(B | A) = 2/3$, and $P(B | \bar{A}) = 3/5$.

$$\begin{aligned}
 P(B) &= P(A \cap B) + P(\bar{A} \cap B) \\
 &= P(A) \cdot P(B | A) + P(\bar{A}) \cdot P(B | \bar{A}) \\
 &= \frac{1}{2} \cdot \frac{2}{3} + \frac{1}{2} \cdot \frac{3}{5} = \frac{19}{30}.
 \end{aligned}$$

5.20 Let E_1 and E_2 be the events that Plant I and II is selected respectively. Then, the probability of the event E that in a batch of 2, one is the production and the other is the maintenance man is

$$\begin{aligned}
 P(E) &= P(E_1 \cap E) + P(E_2 \cap E) \\
 &= P(E_1) \cdot P(E | E_1) + P(E_2) \cdot P(E | E_2) \\
 &= \frac{1}{2} \cdot \frac{5C_1 \cdot 3C_1}{8C_2} + \frac{1}{2} \cdot \frac{4C_1 \cdot 5C_1}{9C_2} \\
 &= \frac{1}{2} \cdot \frac{15}{28} + \frac{1}{2} \cdot \frac{5}{9} = \frac{275}{504}
 \end{aligned}$$

5.22 Let A = event that the item is acceptable; B_1 and B_2 = events that machine is correctly and incorrectly setup, respectively.

Given, $P(A | B_1) = 0.9$; $P(A | B_2) = 0.4$; $P(B_1) = 0.8$ and $P(B_2) = 0.2$. Then $P(B_1 | A) = 0.9$.

5.23 There are two competitors A and B and the lowest bidder gets the contract.

Value of plan = $300 \times 1,000 = 3,00,000$

Contractor A: $P(\text{Bid} < 300) = 0.3$;

$P(\text{Bid} \geq 300) = 0.7$

Contractor B: $P(\text{Bid} < 300) = 0.7$;

$P(\text{Bid} \geq 300) = 0.3$

(i) If both bids are less than Rs 300, probability is $0.3 \times 0.7 = 0.21$. Therefore plan value is: $3,00,000 \times 0.21 = 63,000$.

(ii) If A bids less than 300 and B bids more than 300, probability is $0.3 \times 0.3 = 0.9$. Therefore, plan value is: $3,00,000 \times 0.09 = 27,000$.

(iii) B bids less than 300 while A bids more than 300, probability is: $0.7 \times 0.7 = 0.49$. Therefore plan value is: $3,00,000 \times 0.49 = 1,47,000$.

Therefore, expected value of plan is

$$63,000 + 27,000 + 1,47,000 = 2,37,000.$$

5.24 $P(\text{price of a certain stock not going up}) = 2/3$

$P(\text{price of a certain stock remaining same}) = 1/4$

The probability that the price of the stock will go down during the next week

$$\begin{aligned}
 &= P(\text{price of the stock not going up and not remaining same}) \\
 &= P(\text{price of the stock not going up}) \times P(\text{price of the stock not remaining same}) \\
 &= \left(\frac{2}{3} \right) \times \left(1 - \frac{1}{4} \right) = \left(\frac{2}{3} \right) \times \left(\frac{3}{4} \right) = \frac{1}{2} = 0.5
 \end{aligned}$$

5.25 The assembled part will be defective if any of the parts is defective.

The probability of the assembled part being defective:

$$\begin{aligned}
 &= P[\text{Any of the part is defective}] \\
 &= P[A \cup B] = P(A) + P(B) - P(AB) \\
 &= \frac{9}{100} + \frac{5}{100} - \left(\frac{9}{100} \right) \times \left(\frac{5}{100} \right) = 0.1355
 \end{aligned}$$

The probability that assembled part is not defective

$$= 1 - 0.1355 = 0.8645.$$

5.26 Let A, B, and C denote the respective probabilities of components X, Y, and Z being defective.

$$P(A) = 0.01, P(B) = 0.02, P(C) = 0.05$$

$$\begin{aligned}
 P(A \text{ or } B \text{ or } C) &= P(A) + P(B) + P(C) - P(AB) - P(BC) \\
 &\quad - P(AC) + P(ABC) \\
 &= 0.01 + 0.02 + 0.05 - 0.0002 - 0.0010 \\
 &\quad - 0.0005 + 0.00001 = 0.0784
 \end{aligned}$$

Hence the probability that the assembled product will not be defective = $1 - 0.0784 = 0.9216$.

5.27 Let A be the event that no defective item is produced during a day. Then

$$P(A) = P(I) \cdot P(A|I) + P(2) \cdot P(A|2) + P(3) \cdot P(A|3)$$

The probability that a defective item is produced = 0.02. Probability that a non-defective item is produced = $1 - 0.02 = 0.98$. Also defectives are assumed to occur independently, therefore:

$$P(A | I) = 0.98, P(A | 2) = (0.98)(0.98) \text{ and}$$

$$P(A | 3) = (0.98)(0.98)(0.98)$$

$$\begin{aligned}
 P(A) &= (0.20)(0.98) + (0.35)(0.98)^2 + (0.45)(0.98)^3 \\
 &= 0.1960 + 0.3361 + 0.4322 = 0.9643
 \end{aligned}$$

Hence the probability of no defectives during a day's production is 0.9643.

5.28 A: an engineer has a bachelor's degree only

B: an engineer has a master's degree

C: an engineer is under 30 years of age

D: an engineer is over 40 years of age

$$(a) P(A) = 150/200 = 0.75$$

$$(b) P(B|D) = \frac{P(B \cap D)}{P(D)} = \frac{10/200}{50/200} = 0.20$$

$$(c) P(C|A) = \frac{P(C \cap A)}{P(A)} = \frac{90/200}{150/200} = 0.60$$

5.29 Given that

	<i>Employed</i>	<i>Unemployed</i>	<i>Total</i>
Males	0.40	0.10	0.50
Females	0.475	0.025	0.50
Total	0.875	0.125	1.00

Let M and F be the male and female chosen, respectively.
 $U = \text{Male}$, female chosen is unemployed

$$(a) P(M|U) = \frac{P(M \cap U)}{P(U)} = \frac{0.10}{0.125} = 0.80$$

$$(b) P(F|U) = \frac{P(F \cap U)}{P(U)} = 0.20$$

5.30 The probability that the officer is happy and accedes to requests = 0.6×0.4 .

The probability that the officer is unhappy and accedes to requests = $0.4 \times 0.1 = 0.04$.

Total probability of acceding to requests = $0.24 + 0.04 = 0.28$.

The probability of his being happy if he accedes to requests = $0.24/0.28 = 0.875$.

5.31 (a) $P(D) = P(A) + P(B) + P(C) = 0.35 + 0.08 + 0.01 = 0.44$

$$(b) P(A|D) = \frac{P(A \cap D)}{P(D)} = \frac{0.35}{0.44} = 0.80$$

(c) Since $P(A|D) = 0.80$ and $P(A) = 0.35 + 0.25 = 0.80$, events A and D must be independent.

5.32 Let R = Red toy is chosen and G = Green toy is chosen.

$$\begin{aligned} P(\text{Both toys are } R) &= P(R \text{ on first choice} \cap R \text{ on second choice}) \\ &= P(R \text{ on first choice}) \cdot P(R \text{ on second choice} | R \text{ on first choice}) \\ &= (2/8)(1/7) = 1/28. \end{aligned}$$

5.33 Let A : Executive who would remain with the company despite an equal or slightly better offer

B : Executive who has more than 10 years of service with the company

$$P(A \text{ and } B) = P(A) P(B|A) = (120/200)(75/120) = 0.375$$

5.7 BAYES' THEOREM

In the 18th century, reverend Thomas Bayes, an English Presbyterian minister, raised a question: Does God really exist? To answer this question, he attempted to develop a formula to determine the probability that God does exist, based on evidence that was available to him on earth. Later, Laplace refined Bayes' work and gave it the name *Bayes' Theorem*.

The **Bayes' theorem** is useful in revising the original probability estimates of known outcomes as we gain additional information about these outcomes. The prior probabilities, when changed in the light of new information, are called *revised* or *posterior probabilities*.

Suppose A_1, A_2, \dots, A_n represent n mutually exclusive and collectively exhaustive events with prior marginal probabilities $P(A_1), P(A_2), \dots, P(A_n)$. Let B be an arbitrary event with $P(B) \neq 0$ for which conditional probabilities $P(B|A_1), P(B|A_2), \dots, P(B|A_n)$ are also known. Given the information that outcome B has occurred, the revised (or posterior) probabilities $P(A_i|B)$ are determined with the help of Bayes' theorem using the formula:

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} \quad (5-8)$$

where the posterior probability of events A_i given event B is the conditional probability $P(A_i|B)$.

Since events A_1, A_2, \dots, A_n are mutually exclusive and collectively exhaustive, the event B is bound to occur with either A_1, A_2, \dots, A_n . That is,

$$B = (A_1 \cap B) \cup (A_2 \cap B) \cup \dots \cup (A_n \cap B)$$

where the **posterior probability** of A_i given B is the conditional probability $P(A_i|B)$.

Since $(A_1 \cap B), (A_2 \cap B) \dots (A_n \cap B)$ are mutually exclusive, we get

$$\begin{aligned} P(B) &= P(A_1 \cap B) + P(A_2 \cap B) + \dots + P(A_n \cap B) = \sum_{i=1}^n P(A_i \cap B) \\ &= P(A_1) P(B|A_1) + P(A_2) P(B|A_2) + \dots + P(A_n) P(B|A_n) \\ &= \sum_{i=1}^n P(A_i) P(B|A_i) \end{aligned}$$

Bayes' theorem: A method to compute posterior probabilities (conditional probabilities under statistical dependence).

Posterior probability: A revised probability of an event obtained after getting additional information.

From formula (5-8) for a fixed i , we have

$$\begin{aligned} P(A_i | B) &= \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{P(B)} \\ &= \frac{P(B|A_i) \cdot P(A_i)}{P(A_1) \cdot P(B|A_1) + P(A_2) \cdot P(B|A_2) + \dots + P(A_n) \cdot P(B|A_n)} \end{aligned}$$

Example 5.25: Suppose an item is manufactured by three machines X, Y, and Z. All the three machines have equal capacity and are operated at the same rate. It is known that the percentages of defective items produced by X, Y, and Z are 2, 7, and 12 per cent, respectively. All the items produced by X, Y, and Z are put into one bin. From this bin, one item is drawn at random and is found to be defective. What is the probability that this item was produced on Y?

Solution: Let A be the defective item. We know the prior probability of defective items produced on X, Y, and Z, that is, $P(X) = 1/3$; $P(Y) = 1/3$ and $P(Z) = 1/3$. We also know that

$$P(A|X) = 0.02, \quad P(A|Y) = 0.07, \quad P(A|Z) = 0.12$$

Now, having known that the item drawn is defective, we want to know the probability that it was produced by Y. That is

$$\begin{aligned} P(Y|A) &= \frac{P(A|Y) \cdot P(Y)}{P(X) \cdot P(A|X) + P(Y) \cdot P(A|Y) + P(Z) \cdot P(A|Z)} \\ &= \frac{(0.07) \cdot (1/3)}{(1/3)(0.02) + (1/3)(0.07) + (1/3)(0.12)} = 0.33 \end{aligned}$$

Example 5.26: Assume that a factory has two machines. Past records show that machine 1 produces 30 per cent of the items of output and machine 2 produces 70 per cent of the items. Further, 5 per cent of the items produced by machine 1 were defective and only 1 per cent produced by machine 2 were defective. If a defective item is drawn at random, what is the probability that the defective item was produced by machine 1 or machine 2?

Solution: Let A_1 = Event of drawing an item produced by machine 1,
 A_2 = Event of drawing an item produced by machine 2,
and D = Event of drawing a defective item produced either by machine 1 or machine 2.

From the data in the problem, we know that

$$P(A_1) = 0.30, \quad P(A_2) = 0.70; \quad P(D | A_1) = 0.05, \quad P(D | A_2) = 0.1$$

The data of the problem can now be summarized as under:

Event	Prior Probability $P(A_i)$	Conditional Probability Event $P(D A_i)$	Joint Probability $P(A_i \text{ and } D)$	Posterior (revised) Probability $P(A_i D) P(A_i \text{ and } D)$
(1)	(2)	(3)	(2) \times (3)	
A_1	0.30	0.05	0.015	0.015/0.022 = 0.682
A_2	0.70	0.01	0.007	0.007/0.022 = 0.318

$$\text{Here } P(D) = \sum_{i=1}^2 P(D|A_i) P(A_i) = 0.05 \times 0.30 + 0.01 \times 0.70 = 0.22$$

From the above table, the probability that the defective item was produced by machine 1 is 0.682 or 68.2 per cent and that by machine 2 is only 0.318 or 31.8 per cent. We may now say that the defective item is more likely drawn from the output produced by machine 1.

Example 5.27: A company uses a ‘selling aptitude test’ in the selection of salesmen. Past experience has shown that only 70 per cent of all persons applying for a sales position achieved a classification ‘dissatisfactory’ in actual selling, whereas the remainder were classified as ‘satisfactory’, 85 per cent had scored a passing grade in the aptitude test. Only 25 per cent of those classified dissatisfactory, had passed the test on the basis of this information. What is the probability that a candidate would be a satisfactory salesman given that he passed the aptitude test?

Solution: Let A and B be the event representing ‘unsatisfactory’ classification as a salesman and ‘passing the test’, respectively. Now, the probability that a candidate would be ‘satisfactory’ salesman given that he passed the aptitude test is:

$$P(A|B) = \frac{(0.70)(0.85)}{(0.70)(0.85) + (0.30)(0.25)} = \frac{0.595}{0.595 + 0.075} = 0.888$$

Assuming no change in the type of candidates applying for the selling positions, the probability that a random applicant would be satisfactory is 70 per cent. On the other hand, if the company only accepts an applicant if he passed the test, the probability increases to 88.8 per cent.

Example 5.28: In a bolt factory, machines A, B, and C manufacture 25 per cent, 35 per cent and 40 per cent of the total output respectively. Of the total of their output, 5, 4, and 2 per cent are defective bolts. A bolt is drawn at random and is found to be defective. What is the probability that it was manufactured by machines A, B, or C?

[Punjab Univ., MCom; Madurai Univ., MCom, 1998]

Solution: Let, A_i ($i = 1, 2, 3$) be the event of drawing a bolt produced by machine A, B, and C, respectively. From the data we know that

$$P(A_1) = 0.25; P(A_2) = 0.35, \text{ and } P(A_3) = 0.40$$

From the additional information, we know that

B = the event of drawing a defective bolt

Thus, $P(B|A_1) = 0.05$; $P(B|A_2) = 0.04$; and $P(B|A_3) = 0.02$

The table of posterior probabilities can be prepared as under:

Event	Prior Probability $P(A_i)$	Conditional Probability $P(B A_i)$	Joint Probability (2) \times (3)	Posterior Probability
(1)	(2)	(3)	(4)	(5)
A_1	0.25	0.05	0.0125	$0.0125 \div 0.0345 = 0.362$
A_2	0.35	0.04	0.0140	$0.014 \div 0.0345 = 0.406$
A_3	0.40	0.02	0.0080	$0.008 \div 0.0345 = 0.232$
Total	1.00		0.0345	1.000

The above table shows the probability that the item was defective and produced by machine A is 0.362, by machine B is 0.406, and machine C is 0.232.

Self-Practice Problems 5C

- 5.34** A manufacturing firm produces steel pipes in three plants with daily production volumes of 500, 1000, and 2000 units respectively. According to past experience, it is known that the fractions of defective output produced by the three plants are respectively 0.005, 0.008, and 0.010. If a pipe is selected from a day’s total production and found to be defective, find out (a) from which plant the pipe comes, (b) what is the probability that it came from the first plant? [IIT Roorkee MBA, 2004]

- 5.35** In a post office, three clerks are assigned to process incoming mail. The first clerk, A, processes 40 per cent; the second clerk, B, processes 35 per cent; and the third clerk, C, processes 25 per cent of the mail. The first clerk has an error rate of 0.04, the second has an error rate of 0.06, and the third has an error rate of 0.03. A mail selected at random from a day’s output is found to have an error. The postmaster wishes to know the

probability that it was processed by clerk A or clerk B or clerk C.

- 5.36** A certain production process produces items 10 per cent of which defective. Each item is inspected before supplying to customers but 10 per cent of the time the inspector incorrectly classifies an item. Only items classified as good are supplied. If 820 items have been supplied in all, how many of them are expected to be defective?

- 5.37** A factory produces certain types of output by three machines. The respective daily production figures are: Machine A = 3000 units; Machine B = 2500 units; and Machine C = 4500 units. Past experience shows that 1 per cent of the output produced by machine A is defective. The corresponding fractions of defectives for the other two machines are 1.2 and 2 per cent respectively. An item is drawn at random from the day’s

- production and is found to be defective. What is probability that it comes from the output of (a) Machine A; (b) Machine B; (c) Machine C?
- 5.38** In a bolt factory machines A, B, and C manufacture 25 per cent, 30 per cent and 40 per cent of the total output respectively. Of the total of their output 5, 4, and 2 per cent are defective bolts. A bolt is drawn at random from the lot and is found to be defective. What are the probabilities that it was manufactured by machines A, B, or C?
- 5.39** In a factory manufacturing pens, machines X, Y, and Z manufacture 30, 30, and 40 per cent of the total production of pens, respectively. Of their output 4, 5, and 10 per cent of the pens are defective. If one pen is selected at random, and it is found to be defective, what is the probability that it is manufactured by machine Z?
- 5.40** A worker-operated machine produces a defective item with probability 0.01, if the worker follows the machine's operating instruction exactly, and with probability 0.03 if he does not. If the worker follows the instructions 90 per cent of the time, what proportion of all items produced by the machine will be defective?
- 5.41** Medical case histories indicate that different illnesses may produce identical symptoms. Suppose a particular set of symptoms, 'H' occurs only when one of three illnesses: A, B or C occurs, with probabilities 0.01, 0.005 and 0.02 respectively. The probability of developing the symptoms H, given a illness A, B and C are 0.90, 0.95 and 0.75 respectively. Assuming that an ill person shows the symptoms H, what is the probability that a person has illness A?

Hints and Answers

- 5.34** Let A_1 , A_2 and A_3 = production volume of plant I, II, and III, respectively.

E = defective steel pipe

$$P(A_1) = 500/3500 = 0.1428;$$

$$P(A_2) = 1000/3500 = 0.2857;$$

$$P(A_3) = 2000/3500 = 0.5714$$

$$P(E | A_1) = 0.005, P(E | A_2) = 0.008,$$

$$\text{and } P(E | A_3) = 0.010.$$

$$P(A_1 \cap E) = P(A_1) P(E | A_1) \\ = 0.1428 \times 0.005 = 0.0007;$$

$$P(A_2 \cap E) = P(A_2) P(E | A_2) \\ = 0.2857 \times 0.008 = 0.0022$$

$$P(A_3 \cap E) = P(A_3) P(E | A_3) \\ = 0.5714 \times 0.010 = 0.057$$

$$P(E) = P(A_1 \cap E) + P(A_2 \cap E) + P(A_3 \cap E) \\ = 0.0007 + 0.0022 + 0.057 = 0.0599$$

$$(a) \quad P(A_1 | E) = \frac{P(A_1 \cap E)}{P(E)} = \frac{0.0007}{0.0599} = 0.0116$$

$$P(A_2 | E) = \frac{P(A_2 \cap E)}{P(E)} = \frac{0.0022}{0.0599} = 0.0367;$$

$$P(A_3 | E) = \frac{P(A_3 \cap E)}{P(E)} = \frac{0.057}{0.0599} = 0.951$$

Since $P(A_3 | E)$ is highest, the defective steel pipe has most likely come from the third plant

$$(b) \quad P(A_1 | E) = \frac{P(A_1 \cap E)}{P(E)} = \frac{P(A_1) P(E | A_1)}{P(E)} \\ = \frac{(500/3500) \times 0.005}{0.0599} = 0.0119$$

- 5.35** Let A, B, and C = mail processed by first, second, and third clerk, respectively

E = mail containing error

Given $P(A) = 0.40$, $P(B) = 0.35$, and $P(C) = 0.25$

$$P(E | A) = 0.04, \quad P(E | B) = 0.06,$$

$$\text{and } P(E | C) = 0.03$$

$$\therefore P(A | E) = \frac{P(A) P(E | A)}{P(E)} \\ = \frac{P(A) P(E | A)}{P(A) P(E | A) + P(B) P(E | B) + P(C) P(E | C)} \\ = \frac{0.40 \times 0.04}{0.40 (0.04) + 0.35 (0.06) + 0.25 (0.03)} = 0.36$$

$$\text{Similarly } P(B | E) = [P(B) P(E | B)]/P(E) = 0.47 \\ P(C | E) = [P(C) P(E | C)]/P(E) = 0.17$$

- 5.36** $P(D)$ = Probability of defective item = 0.1; $P(\text{classified as good} | \text{defective}) = 0.1$

$$\therefore P(G) = \text{Probability of good item} = 1 - P(D) \\ = 1 - 0.1 = 0.9$$

$$P(\text{classified as good} | \text{good}) = 1 - P(\text{classified as good} | \text{defective}) \\ = 1 - 0.1 = 0.9$$

$$\therefore P(\text{defective} | \text{classified as good})$$

$$= \frac{P(D) \cdot P(\text{classified as good} | \text{defective})}{[P(D) \cdot P(\text{classified as good} | D) \\ + P(G) P(\text{classified as good} | G)]}$$

$$= \frac{0.1 \times 0.1}{0.1 \times 0.1 + 0.9 \times 0.9} = \frac{0.01}{0.82} = 0.012.$$

- 5.37** (a) 0.20 (b) 0.20 (c) 0.60

- 5.38** $P(A) = 0.37$, $P(B) = 0.40$, $P(C) = 0.23$

- 5.39** $P(Z) = 0.6639$

- 5.40** $P(A) = 0.012$

- 5.41** $P(A | H) = 0.3130$

Formulae Used

1. Counting methods for determining the number of outcomes

- Multiplication method

$$(i) n_1 \times n_2 \times \dots \times n_k$$

$$(ii) n_1 \times n_2 \times \dots \times n_k = n^k$$

- when the event in each trial is the same

- Number of Permutations $n P_r = \frac{n!}{(n-r)!}$

- Number of Combinations $n C_r = \frac{n!}{r!(n-r)!}$

2. Classical or *a priori* approach of computing probability of an event A

$$P(A) = \frac{\text{Number of favourable cases for } A}{\text{All possible cases}} = \frac{c(n)}{c(s)}$$

3. Relative frequency approach of computing probability of an event A in n trials of an experiment

$$P(A) = \lim_{n \rightarrow \infty} \frac{c(A)}{n}$$

4. Rule of addition of two events

- When events A and B are mutually exclusive

$$P(A \text{ or } B) = P(A) + P(B)$$

- When events A and B are not mutually exclusive

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

5. Conditional probability

- For statistically independent events

$$P(A|B) = P(A); P(B|A) = P(B)$$

- For statistically dependent events

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

6. Rule of multiplication of two events

- Joint probability of independent events

$$P(A \text{ and } B) = P(A) \times P(B)$$

- Joint probability of dependent events

$$P(A \text{ and } B) = P(A|B) \times P(B)$$

$$P(A \text{ and } B) = P(B|A) \times P(A)$$

7. Rule of elimination

$$(i) P(B) = \sum P(A_i) P(B|A_i)$$

$$(ii) P(A) = \sum P(B_i) P(A|B_i)$$

8. Baye's rule

$$P(A_i|B) = \frac{P(A_i) P(B|A_i)}{\sum P(A_i) P(B|A_i)}$$

9. Basic rules for assigning probabilities

- The probability assigned to each experimental outcome $0 \leq P(A_i) \leq 1$ for all i

- Sum of the probabilities for all the experimental outcomes

$$\sum P(A_i) = P(A_1) + P(A_2) + \dots + P(A_n) = 1$$

Complement of an event, $P(A) = 1 - P(\bar{A})$

Review Self-Practice Problems

- 5.42** Suppose a nationwide screening programme instituted through schools is being considered to uncover child abuse. It is estimated that 2 per cent of all children are subject to abuse. Further, existing screening programmes are able to determine correctly that abuse occurs 92 per cent of the time and that abuse is incorrectly suspected 5 per cent of the time.

- What is the probability that the results of screening indicating abuse are associated with children who are actually not abused?
- Based upon every 1,00,000 children screened, how many screenings can be expected to lead to a false accusation of abuse?
- Based upon your answer to part (a), is it valid to conclude that 73 per cent of the families not abusing children would be falsely accused? Why or why not?

[Delhi Univ., MBA, 1998]

- 5.43** If there is an increase in capital investment next year, the probability that the price of structural steel will increase is 0.90. If there is no increase in such investment, the probability of an increase is 0.40. Overall, we estimate

that there is a 60 per cent chance that capital investment will increase next year.

- What is the overall probability of an increase in structural steel prices next year?
- Suppose that during the next year structural steel prices in fact increase, what is the probability that there was an increase in capital investment?

[Delhi Univ., MBA, 2000]

- 5.44** A product is assembled from three components X, Y, and Z, and the probability of these components being defective is 0.01, 0.02, and 0.05. What is the probability that the assembled product will not be defective?

[Delhi Univ., MBA, 2001]

- 5.45** A human resource manager has found it useful to categorize engineering job applicants according to their degree in engineering and relevant work experience. Out of all applicants for the job 70 per cent have a degree with or without any work experience, and 60 per cent have work experience with or without the degree. Fifty per cent of the applicants have both the degree and relevant work experience.

- (a) Determine the probability that a randomly selected job applicant has either the degree or relevant work experience.
- (b) What is the probability that the applicant has neither the degree nor work experience?
- 5.46** A salesman is found to complete a sale with 10 per cent of potential customers contacted. If the salesman randomly selects two potential customers and calls on them, then (a) what is the probability that both the calls will result in sales? and (b) what is the probability that the two calls will result in exactly one sale?
- 5.47** Suppose 80 per cent of the material received from a vendor is of exceptional quality, while only 50 per cent of the material received from vendor B is of exceptional quality. However, the manufacturing capacity of vendor A is limited, and for this reason only 40 per cent of the material purchased comes from vendor A. The other 60 per cent comes from vendor B. An incoming shipment of material is inspected and it is found to be of exceptional quality. What is the probability that it came from vendor A.
- 5.48** The municipal corporation routinely conducts two independent inspections of each restaurant, with the restaurant passing only if both inspectors pass it. Inspector A is very experienced, and hence, passes only 2 per cent of restaurants that actually do have rules violations. Inspector B is less experienced and passes 7 per cent restaurants with violations. What is the probability that:
- (a) A reports favourable, given that B has found a violation?
 - (b) B reports favourable with a violation, given that inspector A passes it?
 - (c) A restaurant with a violation is cleared by the corporation.
- 5.49** If a hurricane forms in the Indian Ocean, there is a 76 per cent chance that it will strike the western coast of India. From data gathered over the past 50 years, it has been determined that the probability of a hurricane's occurring in this area in any given year is 0.85. What is the probability that a hurricane will occur in the eastern Indian Ocean and strike India this year?
- 5.50** A departmental store has been the target of many shoplifters during the past month, but owing to increased security precautions, 250 shoplifters have been caught. Each shoplifter's sex is noted, also noted is whether he/she was a first-time or repeat offender. The data are summarized in the table below:
- | Sex | First-Time Offender | Repeat Offender |
|--------|---------------------|-----------------|
| Male | 60 | 70 |
| Female | 44 | 76 |

Assuming that an apprehended shoplifter is chosen at random, find:

- (a) The probability that the shoplifter is male.
- (b) The probability that the shoplifter is a first-time offender, given that the shoplifter is male.
- (c) The probability that the shoplifter is female, given that the shoplifter is a repeat offender.
- (d) The probability that the shoplifter is female, given that the shoplifter is a first-time offender.

5.51 A doctor has decided to prescribe two new drugs to 200 heart patients in the following manner: 50 get drug A, 50 get drug B, and 100 get both. Drug A reduces the probability of a heart attack by 35 per cent drug B, reduces the probability by 20 per cent, and the two drugs, when taken together, work independently. The 200 patients were chosen so that each has an 80 per cent chance of having a heart attack. If a randomly selected patient has a heart attack, what is the probability that the patient was given both drugs?

5.52 The Deputy Commissioner of Police is trying to decide whether to schedule additional patrol units in two sensitive areas, A and B, in his district. He knows that on any given day during the past year, the probabilities of major crimes and minor crimes being committed in area A were 0.478 and 0.602, respectively, and that the corresponding probabilities in area B were 0.350 and 0.523. Assume that major and minor crimes occur independently of each other and likewise that crimes in the two areas are independent of each other.

- (a) What is the probability that no crime of either type will be committed in the area A on a given day?
- (b) What is the probability that a crime of either type will be committed in the area B on a given day?
- (c) What is the probability that no crime of either type will be committed in either areas on a given day?

5.53 The press-room supervisor for a daily newspaper is asked to find ways to print the paper closer to distribution time, thus giving the editorial staff more leeway for last-minute changes. He has the option of running the presses at 'normal' speed or at 110 per cent of normal—'fast' speed. He estimates that these will run at the higher speed 60 per cent of the time. The roll of paper (the newsprint 'web') is twice as likely to tear at the higher speed which would mean stopping the presses temporarily

- (a) If the web on a randomly-selected printing run has a probability of 0.112 of tearing, what is the probability that the web will not tear at normal speed?
- (b) If the probability of tearing at fast speed is 0.20, what is the probability that a randomly-selected torn web occurred at normal speed?

[*Delhi Univ., MBA, 1999*]

5.54 The result of conducting an examination in two papers, A and B, for 20 candidates were recorded as under: 8 passed in paper A, 7 passed in paper B, 8 failed in both papers. If out of these candidates one is selected at random, find the probability that the candidate (a) passed in both A and B, (b) failed only in A, and (c) failed in A or B.

5.55 When two dice are thrown n number of times, the probability of getting at least one double six is greater than 99 per cent. What is the least numerical value of n .

[*CA, Nov., 1998*]

5.56 It is known from past experience that a football team will play 40 per cent of its matches on artificial turf this season. It is also known that a football player's chances of incurring a knee injury are 50 per cent higher if he is playing an artificial turf instead of grass. Further, if a player's probability of knee injury on artificial turf is 0.42, what is the probability that (a) a randomly selected player incurs a knee injury, and (b) a randomly selected

- player with a knee injury, incurred the injury playing on grass?
- 5.57** In a locality of 5000 people, 1200 are above 30 years of age and 3000 are females. Out of 1200 who were above 30 years of age, 200 are females. A person is chosen at random and you are told that the person is female. What is the probability that she is above 30 years of age? [IGNOU, 1997; Delhi Univ., MBA, 1998, 2001]
- 5.58** Suppose 5 men out of 100 and 25 women out of 1000 are colour blind. A colour blind person is chosen at random. What is the probability of his being male (assuming that male and females are equal in proportion).
- 5.59** An organization dealing with consumer products wants to introduce a new product in the market. Based on its past experience, it has a 65 per cent chance of being successful and 35 per cent of not being successful. In order to help the organization to make a decision on the new product, that is, whether to introduce or not, it decides to get additional information on consumer attitude towards the product. For this purpose, the organization decides on a survey. In the past when a product of this type was successful, surveys yielded favourable indication 85 per cent of the time, whereas unsuccessful products received favourable survey indications 30 per cent of the time. Determine the posterior probability of the product being successful given the survey information. [IGNOU, 1999]
- 5.60** Police Head Quarter classified crime by age (in years) of the criminal and whether the crime is violent or non-violent. A total of 150 crimes were reported in the last month as shown in the table below:
- | Type of crime | Age (in years) | | | Total |
|---------------|----------------|-------|---------|-------|
| | Under 20 | 20–40 | Over 40 | |
| Violent | 27 | 41 | 14 | 82 |
| Non-violent | 12 | 34 | 22 | 68 |
| | 39 | 75 | 36 | 150 |
- (a) What is the probability of selecting a case to analyze and finding the crime was committed by someone less than 40 years old.
- (b) What is the probability of selecting a case that involved a violent crime or an offender less than 20 years old?
- (c) If two crimes are selected for review, then what is the probability that both are violent crimes?
- 5.61** With each purchase of a large pizza at a Pizza shop, the customer receives a coupon that can be scratched to see if a prize will be awarded. The odds of winning a free soft drink are 1 in 10, and the odds of winning a free large pizza are 1 in 50. You plan to eat lunch tomorrow at the shop. What is the probability.
- (a) That you will win either a large pizza or a soft drink?
- (b) That you will not win a prize?
- (c) That you will not win a prize on three consecutive visits to the Pizza shop?
- (d) That you will win at least one prize on one of your next three visits to the Pizza shop?
- 5.62** The boxes of men's shirts were received from the factor. Box 1 contained 25 sport shirts and 15 dress shirts. Box 2 contained 30 sport shirts and 10 dress shirts. One of the boxes was selected at random, and a shirt was chosen at random from that box to be inspected. The shirt was a sport shirt. Given this information, what is the probability that the sport shirt came from box 1?
- 5.63** There are four people being considered for the position of chief executive officer of an Enterprises. Three of the applications are over 60 years of age. Two are female, of which only one is over 60. All four applications are either over 60 years of age or female. What is the probability that a candidate is over 60 and female?
- 5.64** A pharmaceutical company through an advertisement in a magazine, estimates that 1 percent of the subscribers will buy products. They also estimate that 05 percent of nonsubscribers will buy the product and that there is one chance in 20 that a person is a subscriber.
- (a) Find the probability that a randomly selected person will buy the products.
- (b) If a person buys the products what is the probability he subscribes to the magazine?
- (c) If a person does not buy the products what is the probability he subscribes to magazine?

Hints and Answers

5.42 (a) 0.727 (b) 6740

5.43 $R =$ rise in price of structural steel,
 $I =$ capital investment increasing.

$$\begin{aligned} P(R) &= P(I \cap R) \cup P(\bar{I} \cap R) \\ &= P(I)P(R|I) + P(\bar{I})P(R|\bar{I}) \\ &= 0.60 \times 0.90 + 0.40 \times 0.40 = 0.70 \end{aligned}$$

$$\begin{aligned} P(I|R) &= \frac{P(I \cap R)}{P(R)} = \frac{P(I)P(R|I)}{P(I)P(R|I) + P(\bar{I})P(R|\bar{I})} \\ &= \frac{0.60 \times 0.90}{0.60 \times 0.90 + 0.40 \times 0.40} \end{aligned}$$

$$= \frac{0.54}{0.70} = 0.77$$

5.44 $P(\text{product not defective})$

$$\begin{aligned} &= P(\bar{X})P(Y)P(Z) + P(X)P(\bar{Y})P(Z) + P(X)P(Y)P(\bar{Z}) \\ &= 0.99 \times 0.02 \times 0.05 + 0.01 \times 0.98 \times 0.05 \\ &\quad + 0.01 \times 0.02 \times 0.95 \\ &= 0.00099 + 0.00049 + 0.00019 = 0.00167 \end{aligned}$$

5.45 $D =$ degree holders; $W =$ with work experience

$$\begin{aligned} (a) \quad P(D \cup W) &= P(D \text{ or } W) \\ &= P(D) + P(W) - P(D \cap W) \\ &= 0.70 + 0.60 - 0.50 = 0.80 \end{aligned}$$

$$(b) P(\bar{D} \cap \bar{W}) = 1.00 - P(D \cup W) \\ = 1.00 - 0.80 = 0.20$$

5.46 S_1, S_2 = calls resulted in sales on both the customers, respectively

$$(a) P(S_1 \text{ and } S_2) = P(S_1 \cap S_2) = P(S_1)P(S_2) \\ = 0.10 \times 0.10 = 0.01$$

$$(b) P(S_1 \cup S_2) = P(S_1 \cap) \cup P(\cap S_2) \\ = 0.10 \times 0.90 + 0.90 \times 0.10 = 0.18$$

5.47 A = material supplied by vendor A

E = material is of exceptional quality.

$$P(A|E) = \frac{P(A \cap E)}{P(E)} = \frac{P(A)P(E|A)}{P(A)P(E|A) + P(B)P(E|B)} \\ = \frac{0.40 \times 0.80}{0.40 \times 0.80 + 0.60 \times 0.50} \\ = \frac{0.32}{0.62} = 0.516$$

5.48 (a) $P(A | \bar{B}) = P(A) = 0.02$

(b) $P(B | A) = P(B) = 0.07$

(c) $P(A \cap B) = P(A)P(B) = 0.02 \times 0.07 = 0.0014$

5.49 Let H = hurricane forming over Indian Ocean;
W = hurricane hits western coast of India,

$$P(H \cap W) = P(H)P(W | H) = 0.76 \times 0.85 = 0.646$$

5.50 M = shoplifter is male, W = shoplifter is female

F = shoplifter is first time offender,

R = shoplifter is repeat offender

(a) $P(M) = (60 + 70) / 250 = 0.520$

(b) $P(F|M) = P(F \cap M)/P(M) = \frac{60}{250} \div \frac{130}{250} = 0.462$

(c) $P(W|R) = P(W \cap R)/P(R) = \frac{76}{250} \div \frac{146}{250} = 0.521$

(d) $P(W|F) = P(W \cap F)/P(F) = \frac{44}{250} \div \frac{104}{250} = 0.423$

5.51 H = heart attack; D = drug given

Drug	$P(D)$	$P(H D)$
A	$50/200 = 0.25$	$0.80 \times 0.65 = 0.520$
B	$50/200 = 0.25$	$0.80 \times 0.80 = 0.640$
A and B	$100/200 = 0.50$	$0.80 \times 0.65 \times 0.80 = 0.416$
$P(H \cap D)$		$P(D H) = P(H \cap D)/P(H)$
0.130		$0.130/0.498 = 0.2610$
0.160		$0.160/0.498 = 0.3213$
0.208		$0.208/0.498 = 0.4177$
$P(H) = 0.498$		

$$P[(A \text{ and } B)/H] = P[(A \text{ and } B) \cap H]/P(H) = 0.208/0.498 = 0.417$$

5.52 M_1, M_2 = major crime in district A and B, respectively
 m_1, m_2 = minor crime in district A and B, respectively.

$$(a) P(M_1 \cup m_1) = P(M_1) + P(m_1) - P(M_1 \cap m_1) \\ = P(M_1) + P(m_1) - P(M_1)P(m_1) \\ = 0.478 + 0.602 - 0.478 \times 0.602 \\ = 0.792$$

$$\therefore P(\bar{M}_1 \cap \bar{m}_1) = 1 - 0.792 = 0.208$$

$$(b) P(M_2 \cup m_2) = P(M_2) + P(m_2) - P(M_2)P(m_2) \\ = 0.350 + 0.523 - 0.350 \times 0.523 \\ = 0.690$$

$$(c) P(\text{crime in A}) = 0.792; P(\text{crime in B}) = 0.690$$

$$P(\text{no crime in A and B}) \\ = 1 - P(\text{crime in at least A or B}) \\ = 1 - [P(A) + P(B) - P(A \text{ and } B)] \\ = 1 - [P(A) + P(B) - P(A)P(B)] = 0.064$$

5.53 (a) Let $x = P(\text{no tear given normal speed})$. Then
 $P(\text{tear}) = P(\text{tear} | \text{normal speed})P(\text{normal speed}) + P(\text{tear} | \text{fast speed})P(\text{fast speed})$
 $0.112 = (1-x)(0.4) + 2(1-x)(0.6) = 1.60 - 1.60x$
 $1.6x = 1.6 - 0.112 = 1.488 \text{ or } x = 1.488/1.6 = 0.93$

(b)

Speed	Prob.	$P(\text{tour} \text{speed})$	$P(\text{tear and speed})$	$P(\text{speed} \text{tear})$
Normal	0.40	0.10	0.04	$0.04/0.16 = 0.25$
Fast	0.60	0.20	0.12	$0.12/0.16 = 0.75$
$P(\text{tear}) = 0.16$				

$$P(\text{normal speed} | \text{tear}) = 0.25$$

5.54 (a) $P(A \cap B) = P(A) + P(B) - P(A \cup B)$

$$\frac{8}{20} + \frac{7}{20} - \left(1 - \frac{8}{20}\right) = \frac{3}{20}$$

(b) $P(\bar{A} \cap B) = P(\bar{A}) \times P(B) = \left(\frac{12}{20}\right)\left(\frac{7}{20}\right) = 0.21$

(c) $P(\bar{A} \cup \bar{B}) = A(\bar{A}) + P(\bar{B}) - (\bar{A} \cap \bar{B}) \\ = \frac{12}{20} + \frac{13}{20} - \frac{8}{20} = \frac{17}{20}$

5.55 Given $1 - (35/36)^n > 0.99$ or $n = 164$.

5.56 A = knee injury; B = playing on artificial turf;
C = playing on grass

(a) $P(A \cap B) = P(B) \cdot P(A | B) = 0.40 \times 0.42 = 0.168$

$P(A \cap C) = P(C)P(A | C) = 0.60 \times 0.28 = 0.168$

Thus $P(A) = P(A \cap B) + P(A \cap C) \\ = 0.168 + 0.168 = 0.336$

(b) $P(C | A) = \frac{P(A \cap C)}{P(A)}$
 $= \frac{P(C) \cdot P(A | C)}{P(C)P(A | C) + P(B)P(A | B)}$
 $= \frac{0.168}{0.168 + 0.168} = \frac{1}{2}$

5.57 $P(A) =$ probability that a person chosen is above 30 years = $1200/5000 = 0.214$

$P(B) =$ probability that a person chosen is female = $3000/5000 = 0.60$

$P(A \text{ and } B) = P(A \cap B) =$ probability that a person chosen is above 30 years and a female

$$= 200/5000 = 0.04$$

But $P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{0.04}{0.06} = \frac{1}{15}$

- 5.58** Let M, F and C denote male, female and colour blind persons. Then

$$\begin{aligned} P(M|C) &= 5/100 = 1/20; \\ P(F|C) &= 25/1000 = 1/40; \\ P(M) &= P(F) = 1/2. \end{aligned}$$

$$\begin{aligned} \text{Thus } P(C|M) &= \frac{P(C \cap M)}{P(M)} \\ &= \frac{P(M) P(M|C)}{P(M) P(M|C) + P(F) P(F|C)} \\ &= \frac{(1/2)(1/20)}{(1/2)(1/20) + (1/2)(1/40)} = \frac{2}{3} \end{aligned}$$

- 5.59** A_1, A_2 = new product is successful and failure, respectively

I = additional information

Event	Probability	Conditional Probability	Joint Probability	Posterior Probability
	(1)	$P(I A_i)$	$P(A_i \cap I)$	$P(A_i I)$
A_1	0.65	0.85	0.552	0.84
A_2	0.35	0.30	0.105	0.16

Posterior probability of product being successful given the survey information is 0.84.

- 5.60** (a) P(crimes both violent and non-violent) committed by a person less than 40 years old)

$$= \frac{39}{150} + \frac{75}{100} = \frac{114}{150} = 0.76$$

- (b) P(crime of violent type or offender less than 20 years old)

$$= \frac{82}{150} + \frac{39}{150} - \frac{27}{150} = \frac{94}{150} = 0.6267$$

- (c) P(both crimes are of violent nature)

$$\frac{82}{150} \times \frac{81}{149} = \frac{6642}{22,300} = 0.2972$$

- 5.61** Let A and B represent the event of winning pizza and soft drink, respectively.

$$\begin{aligned} \text{(a) } P(A \text{ or } B) &= P(A) P(\bar{B}) + P(\bar{A}) P(B) \\ &= \frac{1}{50} \times \frac{9}{10} + \frac{49}{50} \times \frac{1}{10} = \frac{9}{500} + \frac{49}{500} = \frac{58}{500} = 0.116 \end{aligned}$$

$$\begin{aligned} \text{(b) } P(\text{no prize}) &= [1 - P(A)] [1 - P(B)] = P(\bar{A}) P(\bar{B}) \\ &= \frac{49}{500} \times \frac{9}{10} = \frac{441}{500} = 0.882 \end{aligned}$$

$$\begin{aligned} \text{(c) } P(\text{no prize on 3 visits}) &= [P(\text{no prize})]^3 = (0.882)^2 \\ &= 0.686 \end{aligned}$$

$$\begin{aligned} \text{(d) } P(\text{at least one prize}) &= 1 - P(\text{no prize}) = 1 - 0.686 \\ &= 0.314 \end{aligned}$$

- 5.62** Given $P(\text{sport shirt}) = 25/40$, $P(\text{dress shirt}) = 15/40$ (in Box 1); $P(\text{sport short}) = 30/40$; $P(\text{dress short}) = 10/40$ (shirt came from Box 1/shirt was short-shirt)

$$\begin{aligned} &= \frac{P(\text{Box 1 and sport shirt})}{P(\text{sport shirt})} \\ &= \frac{P(\text{sport shirt/Box 1}) P(\text{Box 1})}{P(\text{Box 1}) P(\text{sport/Box 1})} \\ &\quad + P(\text{Box 2}) P(\text{sport shirt/Box 2}) \\ &= \frac{(25/40)(1/2)}{(25/40)(1/2) + (30/40)(1/2)} \\ &= \frac{0.625 \times 0.50}{0.625 \times 0.50 + 0.75 \times 0.50} = \frac{0.3125}{0.6875} = 0.4545 \end{aligned}$$

- 5.63** $P(F \text{ and over 60 years}) = P(F) \times P(\text{over 60 years})$

$$= \frac{2}{4} \times \frac{1}{2} = 0.25$$

- 5.64** Given $P(\text{subscribers buy product}) = 0.05$;
 $P(\text{Non-subscribers buy product}) = 0.95$

$$\begin{aligned} P(\text{buy}|S) &= 0.01, P(\text{not buy}|S) = 0.99; \\ P(\text{buy}|NS) &= 0.005, P(\text{not buy}|NS) = 0.995 \end{aligned}$$

$$\begin{aligned} \text{(a) } P(\text{buy}) &= P(S) P(\text{buy}|S) + P(NS) P(\text{buy}|NS) \\ &= 0.05 \times 0.01 + 0.95 \times 0.005 = 0.00525 \end{aligned}$$

$$\text{(b) } P(S|\text{buy}) = (0.05 \times 0.01)/0.00525 = 0.0952$$

$$\text{(c) } P(S|\text{not buy}) = (0.05 \times 0.99)/0.00525 = 0.0498$$

Case Studies

Case 5.1: Tiger Air Express*

Tiger Air Express Company was founded in 1987, primarily to provide charter air services. Soon after, it incorporated tour business and taxi business into its domain and during its first two years of operations, its sales grew by 168 per cent.

Because of its rapid growth, it experienced some difficulties in providing the necessary resources and building proper infrastructure for this growth. These difficulties were due to underestimated requirement of

capital for vehicles, maintenance, office facilities, and staff. The Gulf War which resulted in steep increase in fuel costs added more burden on the financially-strapped organization.

The management was having difficulty in exactly evaluating the financial and operational situation of the company. A well-known accounting firm was hired to develop and install an accounting system that would accurately reflect the company's financial situation at any

* Adapted from J S Chandan 'Statistics for Business and Economics' *Tiger Air Express Inc.*; Warner Books, 1991, 'What Do Air Shippers Want? Aircargo Survey', *Traffic Management*, July 1992.

given time. As a result of the report, the taxi division was eliminated and the resources were diverted towards the freight and tour divisions of the company, specially the air-freight capabilities.

In order to expand on the air-freight business, further studies were conducted in order to understand the market better. Some insights into the air-freight market were gained through a survey of 44 air freight shippers conducted by an outside agency in 1992. There are two types of services available for forwarding air freight. First, there are integrated carriers. These are the companies which have their own planes. An example is Federal Express. These carriers offer customers pickup and delivery by truck along with airplane shipping. Second, there are freight forwarders. These are the carriers that handle pickup and delivery but use the regular scheduled airlines for air shipment.

A summary of the responses to some of the questions asked in the survey, as noted above, are explained below:

- (i) 30 per cent of the survey respondents stated that their air-freight expenditures had increased in the past year. 60 per cent indicated no change in these expenditures and 10 per cent reported a drop in expenses.
- (ii) 50 per cent of the survey respondents used integrated carriers for shipping. However, for small packages, this percentage increased to 75 per cent. Only 19 per cent used freight forwarders. 62 per cent of those who used freight forwarders did so because of their ability to handle international air freight more efficiently and reliably.

(iii) 39 per cent of the respondents indicated that they were using more 2-3 days deliveries which was cheaper than one day delivery, than they did 2 years ago.

(iv) When the survey asked the customers as to what they looked for in a carrier the responses were as follows:

- On-time delivery : 59 per cent
- Price : 57 per cent
- Customer responsiveness : 43 per cent
- Geographic coverage : 9 per cent
- Single source control : 7 per cent
- Tracing capabilities : 7 per cent.

Questions for Discussion

1. Based on your knowledge of probability theory, what strategy should Tiger Air Express adopt in determining their emphasis on the air freight business, based on the data given?
2. Some of the percentages reported in the survey can be converted into conditional probabilities. Conditional probabilities contain information about the responses in certain categories. Can you describe some of these categories? How would these categories help the management of Tiger Air Express? Explain.
3. If you were a consultant to Tiger Air Express, what recommendations would you give to the management so that they can meet the competition with other companies, offerings based on the type of shippers and shipments that are most probably in demand.

Any body can win unless there happens to be a second entry

—George Alda

When it is not in our power to determine what is true, we ought to follow what is most probable.

— Descartes

Probability Distributions

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- define the terms random variable and probability distribution.
- distinguish between discrete and continuous probability distributions.
- describe the characteristics and compute probabilities using both discrete and continuous probability distributions.
- compute expected value and variance of a random variable.
- apply the concepts of probability distributions to real-life problems.

6.1 INTRODUCTION

In any probabilistic situation each strategy (course of action) may lead to a number of different possible outcomes. For example, a product whose sale is estimated around 100 units, may be equal to 100, less, or more. Here the sale (i.e., an outcome) of the product is measured in real numbers but the volume of the sales is uncertain. The volume of sale which is an uncertain quantity and whose definite value is determined by chance is termed as *random (chance or stochastic) variable*. A listing of all the possible outcomes of a random variable with each outcome's associated probability of occurrence is called *probability distribution*. The numerical value of a random variable depends upon the outcome of an experiment and may be different for different trials of the same experiment. The set of all such values so obtained is called the *range space* of the random variable.

In all such cases as mentioned above, the decision-maker may like to know

- (i) the average value (payoff) of the random variable, and
- (ii) the risk involved in choosing a strategy.

Illustration: If a coin is tossed twice, then the sample space of events, for this random experiment is

$$S = \{H\ H, T\ H, H\ T, T\ T\}$$

In this case, if the decision-maker is interested to know the probability distribution for the number of heads on two tosses of the coin, then a random variable (x) may be defined as:

$$x = \text{number of H's occurred}$$

The values of x will depend on chance and may take the values: H H = 2, H T = 1, T H = 1, T T = 0. Thus the range space of x is {0, 1, 2}.

When a random variable x is defined, a value is given to each simple event in the sample space. The probability of any particular value of x can then be found by adding the probabilities for all the simple events that have that value of x . For example, the probabilities of occurrence of 'heads' can be associated with each of the random variable values. Supposing $P(x = r)$ represents the probability of the random variable taking the value r (here r represents the number of heads occurred). Then probabilities of occurrence of different number of heads are computed as:

Number of Heads (x)	Probability of Outcome $P(x)$
0	$P(x = 0) = P(T T) = P(T) \times P(T) = 0.5 \times 0.5 = 0.25$
1	$P(x = 1) = P(H T) + P(T H) = P(H) \times P(T) + P(T) \times P(H)$ $= 0.5 \times 0.5 + 0.5 \times 0.5 = 0.25 + 0.25 = 0.50$
2	$P(x = 2) = P(H H) = P(H) \times P(H) = 0.5 \times 0.5 = 0.25$

Broad Classes of Random Variable A random variable may be either discrete or continuous. A **discrete random variable** can take on only a finite or countably infinite number of distinct values such as 0, 1, 2, A discrete random variable is usually the result of counting. The number of letters received by a post office during a particular time period, the number of machines breaking down on a given day, the number of vehicles arriving at a toll bridge, and so on, are a few examples of discrete random variables.

A **continuous random variable** can take any numerical value in an interval or collection of intervals. A continuous random variable is usually the result of experimental outcomes that are based on measurement scales. For instance, measurement of time, weight, distance, temperature, and so on are all treated as continuous random variables. Tonnage produced by a steel blast furnace, amount of rainfall in a rainy season, height of individuals, time between arrival of customers at a service system in minutes, are also few examples of continuous random variables.

6.2 PROBABILITY DISTRIBUTION FUNCTION (pdf)

Probability distribution functions can be classified into two categories:

- *Discrete* probability distributions
- *Continuous* probability distributions

A **discrete probability distribution** assumes that the outcomes of a random variable under study can take on *only integer values*, such as:

- A book shop has only 0, 1, 2, ... copies of a particular title of a book
- A consumer can buy 0, 1, 2, ... shirts, pants, etc.

If the random variable x is discrete, its probability distribution called *probability mass function (pmf)* must satisfy following two conditions:

- (i) The probability of a any specific outcome for a discrete random variable must be between 0 and 1. Stated mathematically, $0 \leq f(x=k) \leq 1$, for all value of k
- (ii) The sum of the probabilities over all possible values of a discrete random variable must equal 1. Stated mathematically, $\sum_{\text{all } k} f(x=k) = 1$

A continuous probability distribution assumes that the outcomes of a random variable can take on only value in an interval such as:

- Product costs and prices.
- Floor area of a house, office, etc.

If the random variable x is continuous, then its probability density function must satisfy following two conditions:

$$(i) P(x) \geq 0 ; -\infty < x < \infty \text{ (non-negativity condition)}$$

$$(ii) \int_{-\infty}^{\infty} P(x) dx = 1 \text{ (Area under the continuous curve must total 1)}$$

Continuous probability distribution functions are used to find probabilities associated with random variable values x_1, x_2, \dots, x_n in a given interval or range, say (a, b) . In other words, these probabilities are determined by finding the area under the *pdf* between the values a and b . Mathematically, the area under *pdf* between a and b is given by

$$f(a \leq x \leq b) = f(b) - f(a) = \int_a^b f(x) dx$$

We can express $f(a \leq x \leq b)$ in terms of a distribution function, $F(x)$, provided it is differentiable. That is,

$$\frac{d}{dx} \{f(x)\} = \frac{d}{dx} \left\{ \int_a^b f(x) dx \right\}$$

Illustration: Consider the function, $f(x) = \begin{cases} a & 0 \leq x \leq 5 \\ 0 & \text{otherwise} \end{cases}$

For $f(x)$ to be a *pdf*, the condition, $\int_{-\infty}^{\infty} f(x) dx = 1$ must be satisfied, which

is true if

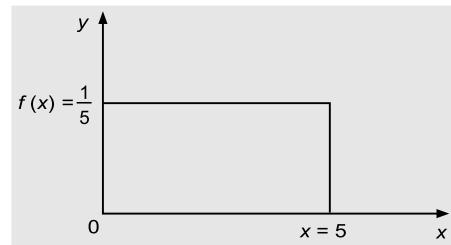
$$\int_0^5 a dx = 1, \text{ i.e. } a = \frac{1}{5}.$$

Since $a > 0$, the function, $f(x) \geq 0$. Thus $f(x)$ satisfies both the conditions for a *pdf*.

Figure 6.1 illustrates the function graphically

Continuous probability distribution: A probability distribution in which the random variable is permitted to take any value within a given range

Figure 6.1
Probability Distribution Function



6.3 CUMULATIVE PROBABILITY DISTRIBUTION FUNCTION (*cdf*)

The cumulative probability distribution function for the continuous random variable x ($-\infty < x < \infty$) is a rule or table that provides the probabilities $P(x \leq k)$ for any real number k . Generally, the term cumulative probability refers to the probability that x is less than or equal to a particular value. For example, if we have three values of a random variable x as: $a < b < c$, then

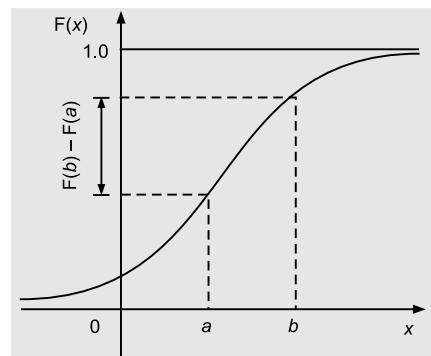
$$\int_a^c f(x) dx \geq \int_a^b f(x) dx$$

This condition shows that *cdf* increases from left to right as shown in Fig. 6.2. Thus the probability that the value of the random variable x is less than any real number a , is given by

$$F(a) = P(x \leq a) = \int_{-\infty}^a f(x) dx$$

where the function $F(a)$, also called the cumulative distribution (or function), represents the probability that x does not exceed a specified value ' a ', and the area under the $f(x)$ curve to the left of the value a . That is, the probability of the random variable x lies at or below some specific value, a . The *cdf* has the properties

Figure 6.2
Cumulative Probability Distribution Function



- (i) $F(a)$ is non-decreasing function
- (ii) $F(-\infty) = 0$ and $F(\infty) = 1$.

In general, given two real numbers a and b such that $a < b$, the probability that the value of x lies in any specified range, say between a and b is,

$$P(a \leq x \leq b) = \int_a^b f(x) dx = \int_{-\infty}^b f(x) dx - \int_{-\infty}^a f(x) dx = F(b) - F(a)$$

A typical *cdf* is illustrated in Fig. 6.2. However, if $P(x = a)$ and $P(x = b)$, then both of these have zero value in a continuous distribution. That is

$$P(x = a) = \int_a^a f(x) dx = 0$$

This result does not mean that the value of the random variable cannot be exactly equal to a . There is an infinitely large number of possible values and the probability associated with any one of them is zero. Thus *cdf* has the following properties:

$$\lim_{a \rightarrow \infty} F(a) = \lim_{a \rightarrow \infty} \int_{-\infty}^a f(x) dx = 1$$

$$\lim_{a \rightarrow -\infty} F(a) = \lim_{a \rightarrow -\infty} \int_{-\infty}^a f(x) dx = 0$$

From this relationship it follows that, $f(x) = \frac{d}{dx} F(x)$.

Figure 6.3

Cumulative Probability Distribution Function

Illustration: For the continuous *pdf* is defined as

$$f(x) = \begin{cases} 1/5, & 0 \leq x \leq 5 \\ 0, & \text{otherwise} \end{cases}$$

the *cdf* in the range $0 \leq x \leq 5$ is given by

$$F(x) = \int_0^x f(x) dx = \int_0^x \frac{1}{5} dx = \frac{x}{5}$$

The *cdf* is illustrated in Fig 6.3.

For a given *pdf*, suppose we want to calculate $P(1 \leq x \leq 3)$. Then

$$P(1 \leq x \leq 3) = F(3) - F(1) = \frac{3}{5} - \frac{1}{5} = \frac{2}{5}$$

This may be seen from Fig. 6.3.

The cumulative probability distribution function of a discrete variable also specifies the probability that an observed value of discrete random variable x will be no greater than a value, say k . In other words, if $F(k)$ is a *cdf* and $f(k)$ is a *pmf*, then

$$F(k) = P(x \leq k) = f(x \leq k)$$

Illustration: Let the probability distribution function of the discrete variable x be as follows:

Random variable :	0	1	2	3	4
Probability, $P(x = a)$:	0.10	0.20	0.40	0.20	0.10

Suppose we want to know the probability of x being equal to or less than 2, that is, $P(x \leq 2) = 0.70$. The probability distribution function and cumulative probability distribution function of 'less than or equal to' type are shown in Table 6.1 and graphed in Fig. 6.4.

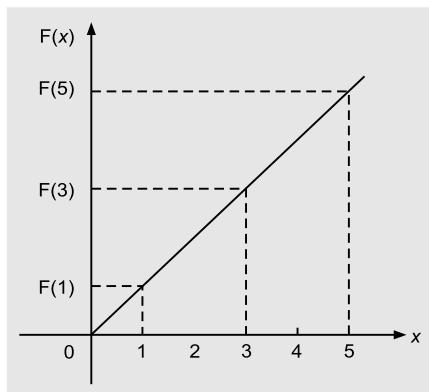
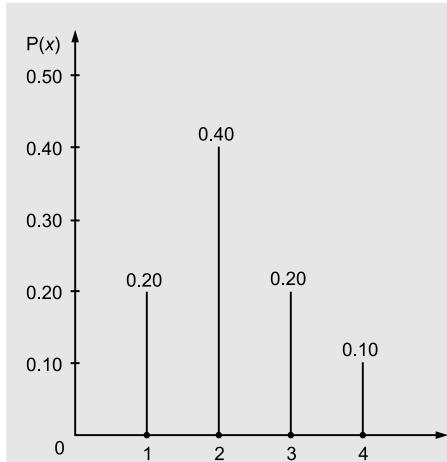
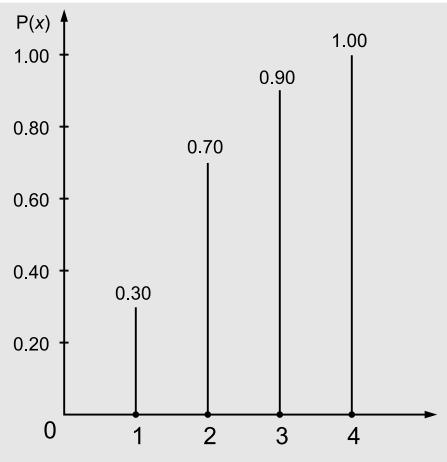


Table 6.1: Cumulative Distribution Function

Random Variable (x)	$P(x = 2)$	$P(x \leq 2)$
0	0.10	0.10
1	0.20	0.30
2	0.40	0.70
3	0.20	0.90
4	0.10	1.00

**Figure 6.4(a) Probability Distribution Function****Figure 6.4(b) Cumulative Distribution Function**

For example, if we want to know the probability of x being greater than 2, then it is given by

$$P(x > 2) = 1 - P(x \leq 2) = 1 - 0.70 = 0.30$$

Thus given the probability distribution function, we can obtain the cumulative distribution function.

6.4 EXPECTED VALUE AND VARIANCE OF A RANDOM VARIABLE

In the same way as discussed in Chapter 3 and 4, a probability distribution is also summarized by its mean and variances.

6.4.1 Expected Value

The mean (also referred as **expected value**) of a random variable is a typical value used to summarize a probability distribution. It is the weighted average, where the possible values of random variable are weighted by the corresponding probabilities of occurrence. If x is a random variable with possible values x_1, x_2, \dots, x_n occurring with probabilities $P(x_1), P(x_2), \dots, P(x_n)$, then the expected value of x denoted by $E(x)$ or μ is the sum of the values of the random variable weighted by the probability that the random variable takes on that value.

$$E(x) = \sum_{j=1}^n x_j P(x_j), \text{ provided } \sum_{j=1}^n P(x_j) = 1$$

Similarly, for the continuous random variable, the expected value is given by:

$$E(x) = \int_{-\infty}^{\infty} x f(x) dx$$

where $f(x)$ is the probability distribution function.

Expected value of a random variable: A weighted average obtained by multiplying each possible value of the random variable with its probability of occurrence.

If $E(x)$ is calculated in terms of rupees, then it is known as *expected monetary value* (EMV). For example, consider the price range of an item along with the probabilities as below:

Price, x	:	50	60	70	80
Probability, $P(x)$:	0.2	0.5	0.2	0.1

Thus the expected monetary value of the item is given by

$$\begin{aligned} \text{EMV}(x) &= \sum_{j=1}^n x_j P(x_j) \\ &= 50 \times 0.2 + 60 \times 0.5 + 70 \times 0.2 + 80 \times 0.1 = \text{Rs } 62. \end{aligned}$$

6.4.2 Variance and Standard Deviation

The expected value measures the *central tendency* of a probability distribution, while variance determines the *dispersion* or *variability* to which the possible random values differ among themselves.

The variance, denoted by $\text{Var}(x)$ or σ^2 of a random variable x is the squared deviation of the individual values from their expected value or mean. That is

$$\begin{aligned} \text{Var}(x) &= E[(x - \mu)^2] = E(x_j - \mu)^2 P(x_j), \text{ for all } j \\ &= E[(x^2 - 2x\mu + \mu^2)] \\ &= E(x^2) - 2\mu E(x) + \mu^2 = E(x^2) - \mu^2 \end{aligned}$$

where $E(x^2) = \sum_{j=1}^n x_j^2 P(x_j)$ and $\mu = \sum_{j=1}^n x_j P(x_j)$

The variance has the disadvantage of squaring the unit of measurement. Thus, if a random variable is measured in rupees, the variance will be measured in rupee squared. This shortcoming can be avoided by using *standard deviation* (σ_x) as a measure of dispersion so as to have the same unit of measurement. That is

$$\sigma_x = \sqrt{\text{Var}(x)} = \sqrt{\sum_{j=1}^n (x_j - \mu)^2 P(x_j)}$$

6.4.3 Properties of Expected Value and Variance

The following are the important properties of an expected value of a random variable:

1. The expected value of a constant c is constant. That is, $E(c) = c$, for every constant c .
2. The expected value of the product of a constant c and a random variable x is equal to constant c times the expected value of the random variable. That is, $E(cx) = cE(x)$.
3. The expected value of a linear function of a random variable is same as the linear function of its expectation. That is, $E(a + bx) = a + bE(x)$.
4. The expected value of the product of two independent random variables is equal to the product of their individual expected values. That is, $E(xy) = E(x) E(y)$.
5. The expected value of the sum of the two independent random variables is equal to the sum of their individual expected values. That is, $E(x + y) = E(x) + E(y)$.
6. The variance of the product of a constant and a random variable X is equal to the constant squared times the variance of the random variable X . That is, $\text{Var}(cx) = c^2 \text{Var}(x)$.
7. The variance of the sum (or difference) of two independent random variables equals the sum of their individual variances. That is, $\text{Var}(x \pm y) = \text{Var}(x) \pm \text{Var}(y)$.

Example 6.1: A doctor recommends a patient to take a particular diet for two weeks and there is equal chance for the patient to lose weight between 2 kgs and 4 kgs. What is the average amount the patient is expected to lose on this diet?

Solution: If x is the random variable, then probability density function is defined as:

$$f(x) = \begin{cases} \frac{1}{2}, & 2 < x < 4 \\ 0, & \text{otherwise} \end{cases}$$

The amount the patient is expected to lose on the diet is:

$$E(x) = \int_2^4 x \cdot \frac{1}{2} dx = \left[\frac{x^2}{4} \right]_2^4 = \frac{1}{4} [(4)^2 - (2)^2] = 3 \text{ kg}$$

Example 6.2: From a bag containing 3 red balls and 2 white balls, a man is to draw two balls at random without replacement. He gains Rs 20 for each red ball and Rs 10 for each white one. What is the expectation of his draw.

Solution: Let x be the random variable denoting the number of red and white balls in a draw. Then x can take up the following values.

$$P(x = 2 \text{ red balls}) = \frac{^3C_2}{^5C_2} = \frac{3}{10}$$

$$P(x = 1 \text{ red and } 1 \text{ white ball}) = \frac{^3C_1 \times ^2C_1}{^5C_2} = \frac{3}{5}$$

$$P(x = 2 \text{ white balls}) = \frac{^2C_2}{^5C_2} = \frac{1}{10}$$

Thus, the probability distribution of x is:

Variable	:	2R	1R and 1W	2W
Gain, x	:	40	30	20
Probability, $P(x)$:		3/10	3/5	1/10

$$\begin{aligned} \text{Hence, expected gain is, } E(x) &= 40 \times (3/10) + 30 \times (3/5) + 20 \times (1/10) \\ &= \text{Rs } 32. \end{aligned}$$

Example 6.3: Under an employment promotion programme, it is proposed to allow sale of newspapers inside buses during off-peak hours. The vendor can purchase newspapers at a special concessional rate of Rs 1.25 per copy against the selling price of Rs 1.50. Any unsold copies are, however, a dead loss. A vendor has estimated the following probability distribution for the number of copies demanded.

Number of copies :	15	16	17	18	19	20
Probability :	0.04	0.19	0.33	0.26	0.11	0.07

How many copies should be ordered so that his expected profit will be maximum?

Solution: Profit per copy = Selling price – Purchasing price = 1.50 – 1.25 = Re 0.25. Thus

$$\text{Expected profit} = \text{Number of copies} \times \text{Probability} \times \text{Profit per copy}$$

The calculations of expected profit are shown in the Table 6.2.

Table 6.2: Calculations of Expected Profit

Number of Copies Demanded	Probability	Profit per Copy (Rs)	Expected Profit (Rs)
(1)	(2)	(3)	(4) = (1) × (2) × (3)
15	0.04	0.25	15
16	0.19	0.25	76
17	0.33	0.25	140
18	0.26	0.25	117
19	0.11	0.25	52
20	0.07	0.25	35

The maximum profit of Rs 140 is obtained when he stocks 17 copies of the newspaper.

Example 6.4: In a cricket match played to benefit an ex-player, 10,000 tickets are to be sold at Rs 500. The prize is a Rs 12,000 fridge by lottery. If a person purchases two tickets, what is his expected gain?

Solution: The gain, say x may take one of two values: he will either lose Rs. 1,000 (i.e. gain will be - Rs 1,000) or win Rs $(12,000 - 1,000) = \text{Rs } 11,000$, with probabilities $9,998/10,000$ and $2/10,000$, respectively. The probability distribution for the gain x is given below:

x	$P(x)$
- Rs 1000	$9,998/10,000$
Rs 11000	$2/10,000$

The expected gain will be

$$\begin{aligned}\mu &= E(x) = \sum x P(x) \\ &= -1000 \times (9,998/10,000) + 11000 \times (2/10,000) = -\text{Rs } 997.6\end{aligned}$$

The result implies that if the lottery were repeated an infinitely large number of times, average or expected loss will be Rs 997.6.

Example 6.5: A market researcher at a major automobile company classified households by car ownership. The relative frequencies of households for each category of ownership are shown below:

<i>Number of cars Per House hold</i>	<i>Relative Frequency</i>
0	0.10
1	0.30
2	0.40
3	0.12
4	0.06
5	0.02

Calculate the expected value and standard deviation of the random variable and interpret the result

[Delhi Univ., MBA, 2003]

Solution: The necessary calculations required to calculate expected and standard deviation of a random variable, say x are shown in Table 6.3.

Table 6.3: Calculations of Expected Value and Standard Deviation

<i>Number of Cars Per Households x</i>	<i>Relative Frequency, P(x)</i>	$x \times P(x)$	$x^2 \times P(x)$
0	0.10	0.10	0.00
1	0.30	0.30	0.30
2	0.40	0.80	1.60
3	0.12	0.36	1.08
4	0.06	0.24	0.96
5	0.02	0.10	0.50
		1.80	4.44

Expected value, $\mu = E(x) = \sum x P(x) = 1.80$. This value indicates that there are on an average 1.8 cars per household

$$\text{Variance, } \sigma^2 = \sum x^2 P(x) - [E(x)]^2 = 4.44 - (1.80)^2 = 4.44 - 3.24 = 1.20$$

$$\text{Standard deviation } \sigma = \sqrt{\sigma^2} = \sqrt{1.20} = 1.095 \text{ cars.}$$

Example 6.6: The owner of a 'Pizza Hut' has experienced that he always sells between 12 and 15 of his famous brand 'Extra Large' pizzas per day. He prepares all of them in advance and store them in the refrigerator. Since the ingredients go bad within one day, unsold pizzas are thrown out at the end of each day. The cost of preparing each pizza is

Rs 120 and he sells each one for Rs 170. In addition to the usual cost, it cost him Rs. 50 per pizza that is ordered but can not be delivered due to insufficient stock. If following is the probability distribution of the number of pizzas ordered each day, then how many 'Extra Large' pizza should he stock each day in order to minimize expected loss.

Number of pizza demanded :	12	13	14	15
Probability :	0.40	0.30	0.20	0.10

Solution: The loss matrix for the given question is shown in Table 6.4

Table 6.4: Pizza Ordered

		Pizza Ordered			
		12	13	14	15
Probability → Pizza Stocked ↓	0.40	0.30	0.20	0.10	Expected Loss
	12	—	100	200	300
13	120	—	100	200	88
14	240	120	—	100	142
15	360	240	120	—	240

Since expected loss of Rs 88 is minimum corresponding to a stock level of 13 pizza, the owner should stock 13 'Extra Large' pizzas each day.

Example 6.7: A company introduces a new product in the market and expects to make a profit of Rs 2.5 lakh during the first year if the demand is 'good'; Rs 1.5 lakh if the demand is 'moderate'; and a loss of Rs 1 lakh if the demand is 'poor.' Market research studies indicate that the probabilities for the demand to be good and moderate are 0.2 and 0.5 respectively. Find the company's expected profit and the standard deviation.

Solution: Let x be the random variable representing profit in three types of demand. Thus, x may assume the values:

$$\begin{aligned}x_1 &= \text{Rs 2.5 lakh when demand is good,} \\x_2 &= \text{Rs 1.5 lakh when demand is moderate, and} \\x_3 &= \text{Rs 1 lakh when demand is poor.}\end{aligned}$$

Since these events (demand pattern) are mutually exclusive and exhaustive, therefore

$$P(x_1) + P(x_2) + P(x_3) = 1 \text{ or } 0.2 + 0.5 + P(x_3) = 1 \text{ or } P(x_3) = 0.3$$

Hence, the expected profit is given by

$$E(x) = 2.5 \times 0.2 + 1.5 \times 0.5 + (-1) \times 0.3 = \text{Rs 0.95 lakh}$$

$$\begin{aligned}\text{Also } E(x^2) &= x_1^2 P(x_1) + x_2^2 P(x_2) + x_3^2 P(x_3) \\&= (2.5)^2 \times 0.2 + (1.5)^2 \times 0.5 + (-1)^2 \times 0.3 = \text{Rs 2.675 lakh}\end{aligned}$$

$$\text{Thus } S.D.(x) = \sqrt{\text{Var}(x)} = \sqrt{E(x^2) - [E(x)]^2} = \sqrt{2.675 - (0.95)^2} = \text{Rs 1.331 lakh.}$$

Conceptual Questions 6A

- Define 'random variable'. How do you distinguish between discrete and continuous random variables. Illustrate your answer with suitable examples.
- (a) Define mathematical expectation of a random variable.
(b) Explain what do you mean by the term 'mathematical expectation'. How is it useful for a businessman? Given an example to illustrate its usefulness.
- What is meant by probability distribution of a random variable? Distinguish between probability density function and probability mass function. Illustrate with examples.
- What do you understand by the expected value of a random variable?
- What are the properties of expected value and variance of a random variable?

[Delhi Univ., MBA, 1997]

Self-Practice Problems 6A

- 6.1** Anil company estimates the net profit on a new product it is launching to be Rs 30,00,000 during the first year if it is ‘successful’; Rs 10,00,000 if it is ‘moderately successful’; and a loss of Rs 10,00,000 if it is ‘unsuccessful’. The firm assigns the following probabilities to its first year prospects for the product: Successful : 0.15, moderately successful : 0.25. What are the expected value and standard deviation of first year net profit for this product?

[Delhi Univ., MBA, 2003]

- 6.2** If the probability that the value of a certain stock will remain the same is 0.46, the probability that its value will increase by Re 0.50 or Re 1.00 per share are respectively 0.17 and 0.23, and the probability that its value will decrease Re 0.25 per share is 0.14, what is the expected gain per share?
- 6.3** A box contains 12 items of which 3 are defective. A sample of 3 items is selected at random from this box. If x represents the number of defective items of 3 selected items, describe the random variable x completely and obtain its expectation.
- 6.4** Fifty per cent of all automobile accidents lead to property damage of Rs 100, forty per cent lead to damage of Rs 500, and ten per cent lead to total loss, that is, damage of Rs 1800. If a car has a 5 per cent chance of being in an accident in a year, what is the expected value of the property damage due to that possible accident?
- 6.5** The probability that there is atleast one error in an account statement prepared by A is 0.2 and for B and C it is 0.25 and 0.4 respectively. A, B, and C prepared 10, 16, and 20 statements respectively. Find the expected number of correct statements in all.
- 6.6** A lottery sells 10,000 tickets at Re 1 per ticket, and the prize of Rs 5000 will be given to the winner of the first draw. Suppose you have bought a ticket, how much should you expect to win?
- 6.7** The monthly demand for transistors is known to have the following probability distribution.

Demand (n) :	1	2	3	4	5	6
Probability (P) :	0.10	0.15	0.20	0.25	0.18	0.12

Determine the expected demand for transistors. Also obtain the variance. Suppose the cost (C) of producing ' n ' transistors is given by the relationship, $C = 10,000 + 500n$. Determine the expected cost.

- 6.8** A bakery has the following schedule of daily demand for cakes. Find the expected number of cakes demanded per day.

Number of Cakes Demanded	Probability
0	0.02
1	0.07
2	0.09
3	0.12
4	0.20
5	0.20
6	0.18
7	0.10
8	0.01
9	0.01

- 6.9** A consignment of machine parts is offered to two firms, A and B, for Rs 75,000. The following table shows the probabilities at which firms A and B will be able to sell the consignment at different prices.

Probability	Price (in Rs) at which the Consignment Can be Sold			
	60,000	70,000	80,000	90,000
A	0.40	0.30	0.20	0.10
B	0.10	0.20	0.50	0.20

Which firm, A or B, will be more inclined towards this offer?

- 6.10** An industrial salesman wants to know the average number of units he sells per sales call. He checks his past sales records and comes up with the following probabilities:

Sales (units) :	0	1	2	3	4	5
Probability :	0.15	0.20	0.10	0.05	0.30	0.20

You are expected to help the salesman in his objective.

- 6.11** A survey conducted over the last 25 years indicated that in 10 years the winter was mild, in 8 years it was cold, and in the remaining 7 it was very cold. A company sells 1000 woollen coats in a mild year, 1300 in a cold year, and 2000 in a very cold year. You are required to find the yearly expected profit of the company if a woollen coat costs Rs 173 and it is sold to stores for Rs 248.

Hints and Answers

- 6.1** x : 3 1 -1
 $P(x)$: 0.15 0.25 $1 - 0.15 - 0.25 = 0.60$
 $E(x) = \text{Rs } 0.10 \text{ million}$, $\text{Var}(x) = \text{Rs } 2.19 \text{ million}$, and
 $\sigma_x = \text{Rs } 1.48 \text{ million}$.

- 6.2** Rs. 0.28
6.3 x : 0 1 2 3
 $P(x)$: 27/64 27/64 9/64 1/64;
 $E(x) = 0.75$

- 6.4** $x : 100 \quad 500 \quad 1,800$
 $P(x) : 0.50 \quad 0.40 \quad 0.10$
 $E(x) = \text{Rs } 430; 0.5 E(x) = \text{Rs } 215$
- 6.5** $P(A) = 0.2; P(B) = 0.25; P(C) = 0.4;$ and $P(\bar{A}) = 0.8;$
 $P(\bar{B}) = 0.75; P(\bar{C}) = 0.6$
 $E(x) = x_1 P(\bar{A}) + x_2 P(\bar{B}) + x_3 P(\bar{C}) = 32$
- 6.6** $P(\text{Win}) = \frac{9999}{10,000}$ and $\frac{1}{1000}$
 $E(x) = -1 \times \frac{9999}{10,000} + 4999 \times \frac{1}{1000} = \text{Rs } 3.9991$
- 6.7** Expected demand for transistors, $E(n) = \Sigma np = 3.62$
 $E(C) = (10,000 + 500n) = 10,000 + 50 E(n)$
 $= \text{Rs } 11,810.$

- 6.8** $E(x) = 508.$
6.9 $\text{EMV}(A) = 6 \times 0.4 + 7 \times 0.3 + 8 \times 0.2 + 9 \times 0.1$
 $= \text{Rs } 70,000.$
 $\text{EMV}(B) = 6 \times 0.1 + 7 \times 0.2 + 8 \times 0.5 + 9 \times 0.2$
 $= \text{Rs } 78,000.$

Firm B will be more inclined towards the offer.

- 6.10** $E(x) = 2.75$

<i>State of Nature</i>	<i>Mild</i>	<i>Cold</i>	<i>Very Cold</i>
Prob. $P(x)$	0.40	0.32	0.28
Sale of coat	1000	1300	2000
Profit, x	$1000 \times$	$1300 \times$	$2000 \times$
	(248 – 173)	(248 – 173)	(248 – 173)

$$E(\text{Profit}) = \text{Rs } 1,03,200$$

6.5 DISCRETE PROBABILITY DISTRIBUTIONS

6.5.1 Binomial Probability Distribution

Binomial probability distribution is a widely used probability distribution for a discrete random variable. This distribution describes discrete data resulting from an experiment called a *Bernoulli process* (named after Jacob Bernoulli, 1654–1705, the first of the Bernoulli family of Swiss mathematicians). For each trial of an experiment, *there are only two possible complementary (mutually exclusive) outcomes such as, defective or good, head or tail, zero or one, boy or girl.* In such cases the outcome of interest is referred to as a ‘success’ and the other as a ‘failure’. The term ‘binomial’ literally means two names.

Bernoulli process: It is a process wherein an experiment is performed repeatedly, yielding either a success or a failure in each trial and where there is absolutely no pattern in the occurrence of successes and failures. That is, the occurrence of a success or a failure in a particular trial does not affect, and is not affected by, the outcomes in any previous or subsequent trials. The trials are independent.

Conditions for Binomial Experiment The Bernoulli process involving a series of independent trials, is based on certain conditions as under:

- There are only two mutually exclusive and collective exhaustive outcomes of the random variable and one of them is referred to as a *success* and the other as a *failure*.
- The random experiment is performed under the same conditions for a fixed and finite (also discrete) number of times, say n . Each observation of the random variable in an random experiment is called a *trial*. Each trial generates either a *success* denoted by p or a *failure* denoted by q .
- The outcome (i.e., success or failure) of any trial is not affected by the outcome of any other trial.
- All the observations are assumed to be independent of each other. This means that the probability of outcomes remains constant throughout the process. Thus, the probability of a success, denoted by p , remains constant from trial to trial. The probability of a failure is $q = 1 - p$.

To understand the Bernoulli process, consider the coin tossing problem where 3 coins are tossed. Suppose we are interested to know the probability of two heads. The possible sequence of outcomes involving two heads can be obtained in the following three ways: HHT, HTH, THH.

The probability of each of the above sequences can be found by using the multiplication rule for independent events. Let the probability of a head be p and the probability of tail be q . The probability of each sequence can be written as:

$$ppq \quad pqp \quad qpq$$

Each of these probabilities can be written as p^2q , they are all equal.

Bernoulli process: A process in which each trial has only two possible outcomes, the probability of the outcome at any trial remains fixed over time, and the trials are statistically independent.

Since three sequences correspond to the same event ‘2 heads’, therefore the probability of 2 heads in 3 tosses is obtained by using the addition rule of probabilities for mutually exclusive events. Since the probability of each sequence is same, we can multiply p^2q (probability of one sequence) by 3 (number of possible sequences or orderings of 2 heads). Hence

$$P(2 \text{ heads}) = 3p^2q = {}^3C_2 p^2q$$

Here it may be noted that the possible sequences equals the binomial coefficient ${}^3C_2 = 3$. This coefficient represents the number of ways that three symbols, of which two are alike (i.e., 2H and one T), can be ordered (or arranged). In general, the binomial coefficient nC_r represents the number of ways that n symbols, of which r are alike, can be ordered.

Since events H and T are equally likely and mutually exclusive, therefore $p = 0.5$ and $q = 0.5$ for a toss of the coin. Thus the probability of 2 heads in 3 tosses, is

$$P(x = 2 \text{ heads}) = {}^3C_2 (0.5)^2 (0.5) = 3 (0.25) (0.5) = 0.375$$

Binomial Probability Function In general, for a binomial random variable, x the probability of success (occurrence of desired outcome) r number of times in n independent trials, regardless of their order of occurrence is given by the formula:

$$P(x = r \text{ successes}) = {}^nC_r p^r q^{n-r} = \frac{n!}{r!(n-r)!} p^r q^{n-r}, r = 0, 1, 2, \dots, n \quad (6-1)$$

where n = number of trials (specified in advance) or sample size

p = probability of success

$q = (1 - p)$, probability of failure

x = discrete binomial random variable

r = number of successes in n trials

In formula (6-1), the term $p^r q^{n-r}$ represents the probability of one sequence where r number of events (called successes) occur in n trials in a particular sequence, while the term nC_r represents the number of possible sequences (combinations) of r successes that are possible out of n trials.

Binomial distribution: A discrete probability distribution of outcomes of an experiment known as a Bernoulli process.

Characteristics of the Binomial Distribution The expression (6-1) is known as **binomial distribution** with parameters n and p . Different values of n and p identify different binomial distributions which lead to different probabilities of r -values. The *mean* and *standard deviation* of a binomial distribution are computed in a shortcut manner as follows:

Mean, $\mu = np$,

Standard deviation, $\sigma = \sqrt{npq}$

Knowing the values of first two central moments $\mu_0 = 1$ and $\mu_1 = 1$, other central moments are given by

Second moment, $\mu_2 = npq$

Third moment, $\mu_3 = npq(q - p)$

Fourth moment, $\mu_4 = 3n^2p^2q^2 + npq(1 - 6pq)$

so that $\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{q - p}{\sqrt{npq}}$, where $\beta_1 = \frac{n^2p^2q^2(q - p)^2}{n^3p^3q^3}$

and $\gamma_2 = \beta_2 - 3 = \frac{\mu_4 - 3}{\mu_2^2} = \frac{1 - 6pq}{npq}$, where $\beta_2 = \frac{3n^2p^2q^2 + npq(1 - 6pq)}{n^2p^2q^2}$

For a binomial distribution, *variance < mean*. This distribution is unimodal when np is a whole number, and $\text{mean} = \text{mode} = np$.

A binomial distribution satisfies both the conditions of *pdf*, because

$$P(x = r) \geq 0 \text{ for all } r = 0, 1, 2, \dots, n$$

$$\sum_{r=0}^n P(x = r) = \sum_{r=0}^n [{}^nC_r p^r q^{n-r}] = (p + q)^n = 1$$

Plotting the Binomial Distributions Let us examine graphically the characteristics of binomial distributions when its parameters n and p change.

- (i) Figure 6.5 illustrates the general shape of a family of binomial distributions with constant $n = 5$ and p varies from 0.3 to 0.7.

In the three cases shown in Fig. 6.5, the skewness varies with the value of p . When p is small (i.e. $p < 0.5$), the distribution is skewed to the right. When p and q are equal (i.e. $p = q = 0.5$), the distribution is symmetric. When p is large (i.e. $p > 0.5$), the distribution is skewed to left.

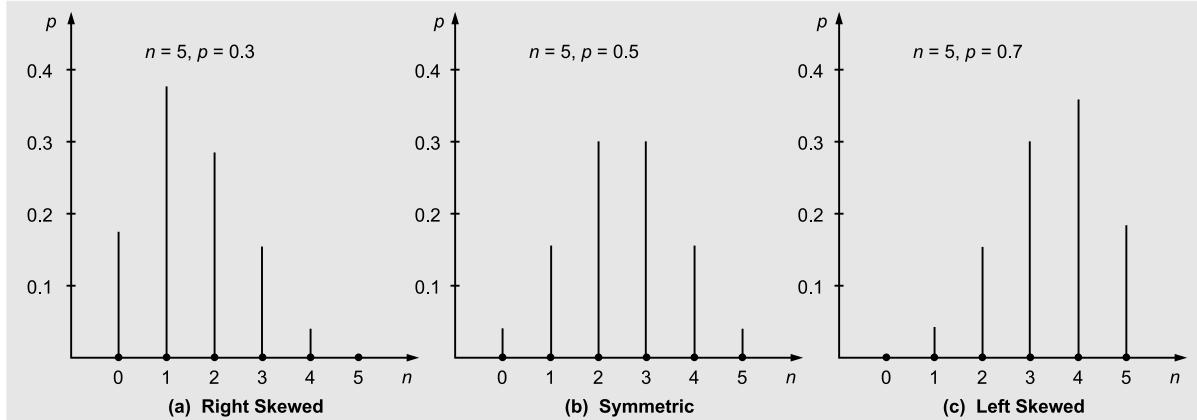


Figure 6.5 Binomial Distributions with Constants n and Variable p

The probability of success is always in the vicinity of the mean, which always increases as p increases. The variance is largest when p and q are equal. The smaller the variance, the larger the probability that a value of random variable, x falls within the vicinity of the mean.

- (ii) If p stays constant but n is increased, then for any value of p other than 0.5 the binomial distribution approaches symmetry. In general, the smaller the value of p , the larger the sample size necessary for the symmetry to occur.

Fitting a Binomial Distribution A binomial distribution can be fitted to the observed values in the data set as follows:

- Find the value of p and q . If one of these is known, the other can be obtained by using the relationship $p + q = 1$.
- Expand $(p + q)^n = p^n + {}^nC_1 p^{n-1} q + {}^nC_2 p^{n-2} q^2 + \dots + {}^nC_r p^{n-r} q^r + \dots + {}^nC_n q^n$ using the concept of binomial theorem.
- Multiply each term in the expansion by the total number of frequencies, N , to obtain the expected frequency for each of the random variable value.

The following recurrence relation can be used for fitting of a binomial distribution:

$$\begin{aligned} f(r) &= {}^nC_r p^r q^{n-r} \\ f(r+1) &= {}^nC_{r+1} p^{r+1} q^{n-r-1} \\ \therefore \frac{f(r+1)}{f(r)} &= \frac{p}{q} \frac{n-r}{r+1} \quad \text{or} \quad f(r+1) = \frac{p}{q} \frac{n-r}{r+1} f(r) \\ \text{For } r = 0, \quad f(1) &= \frac{p}{q} n f(0) \\ \text{For } r = 1, \quad f(2) &= \frac{p}{q} \frac{n-1}{2} f(1) = \left(\frac{p}{q}\right)^2 \frac{n(n-1)}{2!} f(0) \\ \text{For } r = 2, \quad f(3) &= \frac{p}{q} \frac{n-2}{3} f(2) = \left(\frac{p}{q}\right)^3 \frac{n(n-1)(n-2)}{3!} f(0) \end{aligned} \quad (6-2)$$

and so on.

In formula (6-2), we need to calculate $f(0)$, which is equal to q^n , where q can be calculated from the given data.

Example 6.8: A brokerage survey reports that 30 per cent of individual investors have used a discount broker, i.e. one which does not charge the full commission. In a random sample of 9 individuals, what is the probability that

- exactly two of the sampled individuals have used a discount broker?
- not more than three have used a discount broker
- at least three of them have used a discount broker

Solution: The probability that individual investors have used a discount broker is, $p = 0.30$, and therefore $q = 1 - p = 0.70$

(a) Probability that exactly 2 of the 9 individual have used a discount broker is given by

$$\begin{aligned} P(x = 2) &= {}^9C_2 (0.30)^2 (0.70)^7 = \frac{9!}{(9-2)! 2!} (0.30)^2 (0.70)^7 \\ &= \frac{9 \times 8}{2} \times 0.09 \times 0.082 = 0.2656 \end{aligned}$$

(b) Probability that out of 9 randomly selected individuals not more than three have used a discount broker is given by

$$\begin{aligned} P(x \leq 3) &= P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) \\ &= {}^9C_0 (0.30)^0 (0.70)^9 + {}^9C_1 (0.30) (0.70)^8 + {}^9C_2 (0.30)^2 (0.70)^7 \\ &\quad + {}^9C_3 (0.30)^3 (0.70)^6 \\ &= 0.040 + 9 \times 0.30 \times 0.058 + 36 \times 0.09 \times 0.082 \\ &\quad + 84 \times 0.027 \times 0.118 \\ &= 0.040 + 0.157 + 0.266 + 0.268 = 0.731 \end{aligned}$$

(c) Probability that out of 9 randomly selected individuals, at least three have used a discount broker is given by

$$\begin{aligned} P(x \geq 3) &= 1 - P(x < 3) = 1 - [P(x = 0) + P(x = 1) + P(x = 2)] \\ &= 1 - [0.040 + 0.157 + 0.266] = 0.537 \end{aligned}$$

Example 6.9: Mr Gupta applies for a personal loan of Rs 1,50,000 from a nationalised bank to repair his house. The loan offer informed him that over the years, bank has received about 2920 loan applications per year and that the probability of approval was, on average, above 0.85

- Mr Gupta wants to know the average and standard deviation of the number of loans approved per year.
- Suppose bank actually received 2654 loan applications per year with an approval probability of 0.82. What are the mean and standard deviation now?

Solution: (a) Assuming that approvals are independent from loan to loan, and that all loans have the same 0.85 probability of approval. Then

$$\text{Mean, } \mu = np = 2920 \times 0.85 = 2482$$

$$\text{Standard deviation, } \sigma = \sqrt{npq} = \sqrt{2920 \times 0.85 \times 0.15} = 19.295$$

$$(b) \text{ Mean, } \mu = np = 2654 \times 0.82 = 2176.28$$

$$\text{Standard deviation, } \sigma = \sqrt{npq} = \sqrt{2654 \times 0.82 \times 0.18} = 19.792$$

Example 6.10: Suppose 10 per cent of new scooters will require warranty service within the first month of its sale. A scooter manufacturing company sells 1000 scooters in a month,

- Find the mean and standard deviation of scooters that require warranty service
- Calculate the moment coefficient of skewness and kurtosis of the distribution.

Solution: Given that $p = 0.10$, $q = 1 - p = 0.90$ and $n = 1000$

$$(a) \text{ Mean, } \mu = np = 1000 \times 0.10 = 100 \text{ scooters}$$

$$\text{Standard deviation, } \sigma = \sqrt{npq} = \sqrt{1000 \times 0.10 \times 0.90} = 10 \text{ scooters (approx.)}$$

(b) Moment coefficient of skewness

$$\gamma_1 = \sqrt{\beta_1} = \frac{q-p}{\sqrt{npq}} = \frac{0.90-0.10}{\sqrt{9.48}} = 0.084$$

Since γ_1 is more than zero, the distribution is positively skewed.

Moment coefficient of kurtosis, $\gamma_2 = \beta_2 - 3$

$$= \frac{1-6pq}{npq} = \frac{1-6(0.10)(0.90)}{90} = \frac{0.46}{90} = 0.0051$$

Since γ_2 is positive, the distribution is platykurtic.

Example 6.11: The incidence of occupational disease in an industry is such that the workers have 20 per cent chance of suffering from it. What is the probability that out of six workers 4 or more will come in contact of the disease?

[*Lucknow Univ., MBA, 1998; Delhi Univ., MBA, 2002*]

Solution: The probability of a worker suffering from the disease is, $p = 20/100 = 1/5$. Therefore $q = 1-p = 1-(1/5) = 4/5$.

The probability of 4 or more, that is, 4, 5, or 6 coming in contact of the disease is given by

$$\begin{aligned} P(x \geq 4) &= P(x = 4) + P(x = 5) + P(x = 6) \\ &= {}^6C_4 \left(\frac{1}{5}\right)^4 \left(\frac{4}{5}\right)^2 + {}^6C_5 \left(\frac{1}{5}\right)^5 \left(\frac{4}{5}\right) + {}^6C_6 \left(\frac{1}{5}\right)^6 \\ &= \frac{15 \times 16}{15625} + \frac{6 \times 4}{15625} + \frac{1}{15625} = \frac{1}{15625} (240 + 24 + 1) \\ &= \frac{265}{15625} = 0.01695 \end{aligned}$$

Hence the probability that out of 6 workers 4 or more will come in contact of the disease is 0.01695.

Example 6.12: A multiple-choice test contains 8 questions with 3 answers to each question (of which only one is correct). A student answers each question by rolling a balanced dice and checking the first answer if he gets 1 or 2, the second answer if he gets 3 or 4, and the third answer if he gets 5 or 6. To get a distinction, the student must secure at least 75 per cent correct answers. If there is no negative marking, what is the probability that the student secures a distinction?

Solution: Probability of a correct answer, p is one in three so that $p = 1/3$ and probability of wrong answer $q = 2/3$.

The required probability of securing a distinction (i.e., of getting the correct answer of at least 6 of the 8 questions) is given by:

$$\begin{aligned} P(x \geq 6) &= P(x = 6) + P(x = 7) + P(x = 8) \\ &= {}^8C_6 \left(\frac{1}{3}\right)^6 \left(\frac{2}{3}\right)^2 + {}^8C_7 \left(\frac{1}{3}\right)^7 \left(\frac{2}{3}\right) + {}^8C_8 \left(\frac{1}{3}\right)^8 \\ &= \left(\frac{1}{3}\right)^6 \left[{}^8C_6 \left(\frac{2}{3}\right)^2 + {}^8C_7 \left(\frac{1}{3}\right) \left(\frac{2}{3}\right) + {}^8C_8 \left(\frac{1}{3}\right)^2 \right] \\ &= \frac{1}{729} \left[28 \times \frac{4}{9} + 8 \times \frac{2}{9} + \frac{1}{9} \right] = \frac{1}{729} (12.45 + 0.178 + 0.12) \\ &= 0.0196 \end{aligned}$$

Example 6.13: The screws produced by a certain machine were checked by examining the number of defectives in a sample of 12. The following table shows the distribution of 128 samples according to the number of defective items they contained:

No. of defectives	0	1	2	3	4	5	6	7	
in a sample of 12 :	0	1	2	3	4	5	6	7	
No. of samples :	7	6	19	35	30	23	7	1	= 128

- (a) Fit a binomial distribution and find the expected frequencies if the chance of machine being defective is 0.5.
 (b) Find the mean and standard deviation of the fitted distribution.

[Delhi Univ., MBA, 2003]

Solution: (a) The probability of a defective screw is, $p = 1/2$ and therefore $q = 1 - p = 1/2$; $N = 128$.

Since there are 8 terms, therefore $n = 7$. Thus, the probability that the defective items are 0, 1, 2, ..., 7 is given by:

$$(p + q)^n = p^n + {}^nC_1 p^{n-1} q + {}^nC_2 p^{n-2} q^2 + \dots + {}^nC_7 p^{n-7} q^7$$

or

$$\left(\frac{1}{2} + \frac{1}{2}\right)^7 = \left(\frac{1}{2}\right)^7 + {}^7C_1 \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right) + {}^7C_2 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^2 + \dots + {}^7C_7 \left(\frac{1}{2}\right)^7$$

$$= \left(\frac{1}{2}\right)^7 [1 + 7 + 21 + 35 + 35 + 21 + 7 + 1]$$

For obtaining the expected frequencies, multiply each term by $N = 128$. That is,

$$128 \left(\frac{1}{2} + \frac{1}{2}\right)^7 = 128 \times \frac{1}{128} (1 + 7 + 21 + 35 + 35 + 21 + 7 + 1)$$

Thus, the expected frequencies are

$x :$	0	1	2	3	4	5	6	7
$f :$	1	7	21	35	35	21	7	1

(b) The mean of binomial distribution is given by np and standard deviation by \sqrt{npq} . Given that, $n = 7$, $p = q = 1/2$. Thus

$$\text{Mean} = np = 7 \times (1/2) = 3.5$$

$$\text{Standard deviation} = \sqrt{npq} = \sqrt{7 \times (1/2) \times (1/2)} = \sqrt{1.75} = 1.32.$$

Conceptual Questions 6B

6. (a) Define binomial distribution stating its parameters, mean, and standard deviation, and give two examples where such a distribution is ideally suited.
 (b) Define binomial distribution. Point out its chief characteristics and uses. Under what conditions does it tend to Poisson distribution?
7. For a binomial distribution, is it true that the mean is the most likely value? Explain.
8. Demonstrate that the binomial coefficient nC_r equals ${}^nC_{n-r}$ and illustrate this with a specific numerical example.
9. What assumptions must be met for a binomial distribution to be applied to a real life situation?
10. What is meant by the term parameter of a probability distribution? Relate the concept to the binomial distribution?
11. What information is provided by the mean, standard deviation, and central moments of the binomial distribution?
12. What is a binomial coefficient and illustrate this with a specific numerical example.

Self-Practice Problems 6B

- 6.12 The normal rate of infection of a certain disease in animals is known to be 25 per cent. In an experiment with 6 animals injected with a new vaccine it was observed that none of the animals caught the infection. Calculate the probability of the observed result.
- 6.13 Out of 320 families with 5 children each, what percentage would be expected to have (i) 2 boys and 3 girls,

(ii) at least one boy? Assume equal probability for boys and girls.

- 6.14 The incidence of a certain disease is such that on an average 20 per cent of workers suffer from it. If 10 workers are selected at random, find the probability that (i) exactly 2 workers suffer from the disease, (ii) not more than 2 workers suffer from the disease.

Calculate the probability upto fourth decimal place.

[MD Univ., MCom, 1998]

- 6.15** The mean of a binomial distribution is 40 and standard deviation 6. Calculate n , p , and q .

[Delhi Univ., MBA, 1998]

- 6.16** A student obtained answers with mean $\mu = 2.4$ and variance $\sigma^2 = 3.2$ for a certain problem given to him using binomial distribution. Comment on the result.

- 6.17** The probability that an evening college student will graduate is 0.4. Determine the probability that out of 5 students (a) none, (b) one, and (c) at least one will graduate. [Madras Univ., MCom, 1997]

- 6.18** The normal rate of infection of a certain disease in animals is known to be 25 per cent. In an experiment with 6 animals injected with a new vaccine it was observed that none of the animals caught infection. Calculate the probability of the observed result.

- 6.19** Is there any inconsistency in the statement that the mean of a binomial distribution is 20 and its standard deviation is 4? If no inconsistency is found what shall be the values of p , q , and n .

- 6.20** A multi-choice test consists of 8 questions with 3 answers to each question (of which only one is correct). A student answers each question by rolling a balanced dice and selects the first answer if he gets 1 or 2, the second if he

gets 3 or 4, and the third answer if he gets 5 or 6. To get a distribution, the student must secure at least 75 per cent correct answers. If there is no negative marking, what is the probability that the student secures a distinction?

- 6.21** Find the probability that in a family of 5 children there will be (i) at least one boy (ii) at least one boy and one girl (Assume that the probability of a female birth is 0.5).

- 6.22** A famous advertising slogan claims that 4 out of 5 housewives cannot distinguish between two particular brands of butter. If this claim is valid and 5000 housewives are tested in groups of 5, how many of these groups will contain 0, 1, 2, 3, 4, and 5 housewives who do not distinguish between the two products? Assume that the capacity to distinguish between the two brands is randomly distributed so that Bernoulli trial conditions are satisfied.

- 6.23** A supposed coffee connoisseur claims that he can distinguish between a cup of instant coffee and a cup of percolator coffee 75 per cent of the time. It is agreed that his claim will be accepted if he correctly identifies at least 5 out of 6 cups. Find (a) his chance of having the claim accepted if he is in fact only guessing, and (b) his chance of having the claim rejected when he does have the ability he claims.

Hints and Answers

- 6.12** Let P denote infection of the disease. Then

$$p = 25/100 = 1/4 \text{ and } q = 3/4.$$

$$P(x=0) = {}^6C_0 \left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^6 = \frac{729}{4096}$$

- 6.13** (i) Given $p = q = 1/2$

$$\begin{aligned} P(\text{boy} = 2) &= {}^5C_2 p^2 q^3 = {}^5C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3 \\ &= \frac{5}{16} \text{ or } 31.25\% \end{aligned}$$

$$\text{(ii)} \quad P(\text{boy} \geq 1) = 1 - {}^5C_0 q^5 = 1 - \frac{1}{32}$$

$$= \frac{31}{32} \text{ or } 97 \text{ per cent.}$$

- 6.14** Probability that a worker suffers from a disease, $p = 1/5$ and $q = 4/5$.

$$\begin{aligned} P(x=r) &= {}^nC_r q^{n-r} p^r = {}^{10}C_r \left(\frac{1}{5}\right)^r \left(\frac{4}{5}\right)^{10-r} \\ &= {}^{10}C_r \frac{4^{10-r}}{5^{10}}; r = 0, 1, 2, \dots, 10 \end{aligned}$$

$$\text{(i)} \quad P(x=2) = {}^{10}C_2 \frac{4^{10-2}}{5^{10}} = 0.302$$

$$\begin{aligned} \text{(ii)} \quad P(x=0) + P(x=1) + P(x=2) \\ = \frac{1}{5^{10}} ({}^{10}C_0 4^{10} + {}^{10}C_1 4^9 + {}^{10}C_2 4^8) = 0.678. \end{aligned}$$

- 6.15** Given $\mu = np = 40$ and $\sigma = \sqrt{npq} = 6$. Squaring σ , we get $npq = 36$ or $40q = 36$ or $q = 0.9$. Then $p = 1 - q = 0.28$.

Since $np = 40$ or $n = 40/p = 40/0.28 = 143$.

- 6.16** Given $\sigma^2 = npq = 3.2$ and $\mu = np = 2.4$. Then $2.4q = 3.2$ or $q = 3.2/2.4 = 1.33$ (inconsistent result)

- 6.17** Given $p = 0.4$ and $q = 0.6$

$$\begin{aligned} \text{(a)} \quad P(x=\text{no graduate}) &= {}^5C_0 (0.4)(0.6)^5 \\ &= 1 \times 1 \times 0.0777 = 0.0777 \end{aligned}$$

$$\text{(b)} \quad P(x=1) = {}^5C_1 (0.4)^1 (0.6)^4 = 0.2592$$

$$\text{(c)} \quad P(x \geq 1) = 1 - P(x=0) = 1 - 0.0777 = 0.9223$$

- 6.18** Probability of infection of disease = $25/100 = 0.25$; $q = 1 - p = 0.75$.

The first term in the expansion of $(q+p)^n = \left(\frac{3}{4} + \frac{1}{4}\right)^6$ is ${}^6C_0 \left(\frac{3}{4}\right)^6 = 0.177$, which is also the required probability.

- 6.19** Given $\mu = np = 20$ and $\sigma = \sqrt{npq} = 4$ or $npq = 16$ or $20q = 16$ or $q = 16/20 = 0.80$ and then $p = 1 - q = 0.20$. Hence $npq = 16$ gives $n = 16/pq = 16/(0.20 \times 0.80) = 100$.

- 6.20** Probability of correct answer, $p = 1/3$ and wrong answer, $q = 2/3$.

Probability of securing distinction (i.e., answering at least 6 of 8 questions correctly). That is,

$$\begin{aligned} P(x \geq 6) &= P(x = 6) + P(x = 7) + P(x = 8) \\ &= {}^8C_6 \left(\frac{1}{3}\right)^6 \left(\frac{2}{3}\right)^2 + {}^8C_7 \left(\frac{1}{3}\right)^7 \left(\frac{2}{3}\right) + {}^8C_8 \left(\frac{1}{3}\right)^8 \\ &= \left(\frac{1}{3}\right)^6 \left[28 \times \frac{4}{9} + 8 \times \frac{1}{3} \times \frac{2}{3} + \frac{1}{9} \right] = \frac{1}{729} \times \frac{129}{9} \\ &= 0.019 \end{aligned}$$

6.21 Since $p = q = 0.5$, therefore

$$\begin{aligned} \text{(i)} \quad P(\text{boy} = 0) &= {}^5C_0 (0.5)^0 (0.5)^5 = 0.031 \\ P(\text{at least one boy}) &= 1 - 0.031 = 0.969 \\ \text{(ii)} \quad P(\text{at least 1B and 1G}) &= {}^5C_1 (0.5)^1 (0.5)^4 \\ &\quad + {}^5C_2 (0.5)^2 (0.5)^3 + {}^5C_3 (0.5)^3 (0.5)^2 \\ &\quad + {}^5C_4 (0.5)^4 (0.5) \\ &= \frac{5}{32} + \frac{10}{32} + \frac{10}{32} + \frac{5}{32} = \frac{30}{32} \end{aligned}$$

6.22 p = probability that 4 out of 5 cannot distinguish between two brands = $4/5 = 0.8$

so that $q = 1 - p = 1/5 = 0.2$

Expected distribution containing 0, 1, . . . , 5 housewives who do not distinguish between two brands

$$\begin{aligned} &= {}^5C_0 (0.2)^5 + {}^5C_1 (0.8) (0.2)^4 + {}^5C_2 (0.8)^2 (0.2)^3 \\ &\quad + {}^5C_3 (0.8)^3 (0.2)^2 + {}^5C_4 (0.8)^4 (0.2)^1 \\ &\quad + {}^5C_5 (0.8)^5 \end{aligned}$$

Thus required number in each group would be

$$\begin{aligned} &5000 (0.2)^5; 5000 {}^5C_1 (0.8) (0.2)^4; \\ &5000 {}^5C_2 (0.8)^2 (0.2)^3; 5000 {}^5C_3 (0.8)^3 (0.2)^2; \\ &5000 {}^5C_4 (0.8)^4 (0.2); 5000 (0.8)^5 \end{aligned}$$

6.23 Given p = probability that he is capable of making a distinction = 0.75; $q = 1 - p = 0.25$

$$\begin{aligned} \text{(i)} \quad P(x < 5) &= 1 - P(x \geq 5) = 1 - [{}^6C_5 (0.75)^5 (0.25) \\ &\quad + {}^6C_6 (0.75)^6] \\ &= 1 - 0.534 = 0.466 \end{aligned}$$

$$\text{(ii)} \quad P(x \geq 5) = 0.534$$

6.5.2 Poisson Probability Distribution

Poisson distribution is named after the French mathematician S. Poisson (1781–1840), The Poisson process measures the number of occurrences of a particular outcome of a discrete random variable in a *predetermined time interval, space, or volume*, for which an *average number* of occurrences of the outcome is known or can be determined. In the Poisson process, the random variable values need counting. Such a count might be (i) number of telephone calls per hour coming into the switchboard, (ii) number of fatal traffic accidents per week in a city/state, (iii) number of patients arriving at a health centre every hour, (iv) number of organisms per unit volume of some fluid, (v) number of cars waiting for service in a workshop, (vi) number of flaws per unit length of some wire, and so on. The Poisson probability distribution provides a simple, easy-to compute and accurate approximation to a binomial distribution when the probability of success, p is very small and n is large, so that $\mu = np$ is small, preferably $np > 7$. It is often called the '*law of improbable*' events meaning that the probability, p , of a particular event's happening is very small. As mentioned above **Poisson distribution** occurs in business situations in which there are a few successes against a large number of failures or vice-versa (i.e. few successes in an interval) and has single independent events that are mutually exclusive. Because of this, the probability of success, p is very small in relation to the number of trials n , so we consider only the probability of success.

Poisson distribution: A discrete probability distribution in which the probability of occurrence of an outcome within a very small time period is very small, and the probability that two or more such outcomes will occur within the same small time interval is negligible. The occurrence of an outcome within one time period is independent of the other.

Conditions for Poisson Process The use of Poisson distribution to compute the probability of the occurrence of an outcome during a specific time period is based on the following conditions:

- (i) The outcomes within any interval occur randomly and independently of one another.
- (ii) The probability of one occurrence in a small time interval is proportional to the length of the interval and independent of the specific time interval.
- (iii) The probability of more than one occurrence in a small time interval is negligible when compared to the probability of just one occurrence in the same time interval.
- (iv) The average number of occurrences is constant for all time intervals of the same size.

Poisson Probability Function If the probability, p of occurrence of an outcome of interest (i.e., success) in each trial is very small, but the number of independent trials n is sufficiently large, then the average number of times that an event occurs in a certain period of time or space, $\lambda = np$ is also small. Under these conditions the binomial probability function

$$\begin{aligned} P(x = r) &= {}^nC_r p^r q^{n-r} = \frac{n(n-1)(n-2)\dots(n-r+1)}{r!} p^r q^{n-r} \\ &= \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{r-1}{n}\right) \frac{\lambda^r}{r!} \left(1 - \frac{\lambda}{n}\right)^{n-r}; np = \lambda \text{ or } p = \lambda/n \end{aligned}$$

tends to $\frac{\lambda^r}{r!} e^{-\lambda}$ for a fixed r . Thus the Poisson probability distribution which approximates the binomial distribution is defined by the following probability function:

$$P(x = r) = \frac{\lambda^r e^{-\lambda}}{r!}, r = 0, 1, 2, \dots \quad (6-3)$$

where $e = 2.7183$.

Characteristics of Poisson Distribution Since Poisson probability distribution is specified by a process rate λ and the time period t , its mean and variance are identical and are expressed in terms of the parameters: n and p as shown below:

- The arithmetic mean, $\mu = E(x)$ of Poisson distribution is given by

$$\begin{aligned} \mu &= \sum x P(x) = \sum x \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 1, 2, 3, \dots \quad \text{and} \quad x P(x) = 0 \text{ for } x = 0 \\ &= \lambda e^{-\lambda} + \lambda^2 e^{-\lambda} + \frac{\lambda^3 e^{-\lambda}}{2!} + \dots + \frac{\lambda^r e^{-\lambda}}{(x-1)!} + \dots \\ &= \lambda e^{-\lambda} \left[1 + \lambda + \frac{\lambda^2}{2!} + \dots + \frac{\lambda^{r-1}}{(x-1)!} + \dots \right] = \lambda e^{-\lambda} e^\lambda = \lambda \end{aligned}$$

Thus the mean of the distribution is $\mu = \lambda = np$.

- The variance σ^2 of Poisson distribution is given by

$$\begin{aligned} \sigma^2 &= E(x^2) - [E(x)]^2 = E(x^2) - \lambda^2 \\ &= \sum x^2 \frac{e^{-\lambda} \lambda^x}{x!} - \lambda^2 = e^{-\lambda} \sum \frac{x(x-1)+x}{x!} \lambda^x - \lambda^2 \\ &= \lambda^2 e^{-\lambda} \sum \frac{\lambda^{x-2}}{(x-2)!} + \lambda e^{-\lambda} \sum \frac{\lambda^{x-1}}{(x-1)!} - \lambda^2 \\ &= \lambda^2 e^{-\lambda} e^\lambda + \lambda e^{-\lambda} e^\lambda - \lambda^2 = \lambda^2 + \lambda - \lambda^2 = \lambda \end{aligned}$$

Thus the variance of the distribution is $\sigma^2 = \lambda = np$.

The central moments of Poisson distribution can also be determined by the following recursion relation:

$$\mu_r = E(x - \lambda)^r = \sum (x - \lambda)^r e^{-\lambda} \frac{\lambda^x}{x!}$$

Differentially μ_r with respect to λ , we have

$$\frac{d\mu_r}{d\lambda} = -r\mu_{r-1} + \frac{\mu_{r+1}}{\lambda} \quad \text{or} \quad \mu_{r+1} = \lambda \left[r\mu_{r-1} + \frac{d\mu_r}{d\lambda} \right] \quad (6-4)$$

Substituting $\mu_0 = 1$ and $\mu_1 = 0$ and putting $r = 1, 2$ and 3 in Eqn. (6-4), we have

$$\begin{aligned} \mu_2 &= \mu_3 = \lambda \\ \mu_4 &= \lambda + 3\lambda^2 \end{aligned}$$

so that $\gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{\lambda}}$ and $\gamma_2 = \beta_2 - 3 = \frac{1}{\lambda}$

Hence Poisson distribution is defined by the parameter λ and is positively skewed and leptokurtic. This implies that there is a possibility of infinitely large number of occurrences in a particular time interval, even though the average rate of occurrences is very small. However, as $\lambda \rightarrow \infty$, the distribution tends to be symmetrical and mesokurtic.

It is very rare for more than one event to occur during a short interval of time. The shorter the duration of interval, the occurrence of two or more events also becomes rare. The probability that exactly one event will occur in such an interval is approximately λ times its duration.

If λ is not an integer and $m = [\lambda]$, the largest integer contained in it, then m is the unique mode of the distribution. But if λ is an integer, the distribution would be bimodal.

The typical application of Poisson distribution is for analysing queuing (or waiting line) problems in which arriving customers during an interval of time arrive independently and the number of arrivals depends on the length of the time interval. While applying Poisson distribution if we consider a time period of different length, the distribution of number of events remains Poisson with the mean proportional to the length of the time period.

Fitting a Poisson Distribution A Poisson distribution can be fitted to the observed values in the data set by simply obtaining values of λ and calculating the probability of zero occurrence. Other probabilities can be calculated by the recurrence relation as follows:

$$f(r) = \frac{e^{-\lambda} \lambda^r}{r!}$$

$$f(r+1) = \frac{e^{-\lambda} \lambda^{r+1}}{(r+1)!}$$

$$\text{or } \frac{f(r+1)}{f(r)} = \frac{\lambda}{r+1} \quad \text{or } f(r+1) = \frac{\lambda}{(r+1)} \cdot f(r); \quad r = 0, 1, 2, \dots$$

Thus, for $r = 0$, $f(1) = \lambda f(0)$,

$$\text{for } r = 1, \quad f(2) = \frac{\lambda}{2} f(1) = \frac{\lambda^2}{2} f(0)$$

and so on, where $f(0) = e^{-\lambda}$.

After obtaining the probability for each of the random variable values, multiply each of them by N (total frequency) to get the expected frequency for the respective values.

Example 6.14: What probability model is appropriate to describe a situation where 100 misprints are distributed randomly throughout the 100 pages of a book? For this model, what is the probability that a page observed at random will contain at least three misprints?

Solution: Since 100 misprints are distributed randomly throughout the 100 pages of a book, therefore on an average there is only one mistake on a page. This means, the

probability of there being a misprint, $p = 1/100$, is very small and the number of words, n , in 100 pages are very large. Hence, Poisson distribution is best suited in this case.

Average number of misprints in one page, $\lambda = np = 100 \times (1/100) = 1$. Therefore $e^{-\lambda} = e^{-1} = 0.3679$.

Probability of at least three misprints in a page is

$$\begin{aligned} P(x \geq 3) &= 1 - P(x < 3) = 1 - \{P(x = 0) + P(x = 1) + P(x = 2)\} \\ &= 1 - [e^{-\lambda} + \lambda e^{-\lambda} + \frac{1}{2!} \lambda^2 e^{-\lambda}] \\ &= 1 - \left\{ e^{-1} + e^{-1} + \frac{e^{-1}}{2!} \right\} = 1 - 2.5 e^{-1} = 1 - 2.5 (0.3679) \\ &= 0.0802 \end{aligned}$$

Example 6.15: A new automated production process has had an average of 1.5 breakdowns per day. Because of the cost associated with a breakdown, management is concerned about the possibility of having three or more breakdowns during a day. Assume that breakdowns occur randomly, that the probability of a breakdown is the same for any two time intervals of equal length, and that breakdowns in one period are independent of breakdowns in other periods. What is the probability of having three or more breakdowns during a day?
[HP Univ., MBA, 1995; Kumaon Univ., 1998]

Solution: Given that, $\lambda = np = 1.5$ breakdowns per day. Thus probability of having three or more breakdowns during a day is given by

$$\begin{aligned} P(x \geq 3) &= 1 - P(x < 3) = 1 - [P(x=0) + P(x=1) + P(x=2)] \\ &= 1 - \left[\frac{\lambda^0 e^{-\lambda}}{0!} + \frac{\lambda^1 e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} \right] \\ &= 1 - e^{-\lambda} \left[1 + \lambda + \frac{1}{2} \lambda^2 \right] = 1 - 0.2231 \left[1 + 1.5 + \frac{1}{2} (1.5)^2 \right] \\ &= 1 - 0.2231 (3.625) = 1 - 0.8088 = 0.1912 \end{aligned}$$

Example 6.16: Suppose a life insurance company insures the lives of 5000 persons aged 42. If studies show the probability that any 42-years old person will die in a given year to be 0.001, find the probability that the company will have to pay at least two claims during a given year.

Solution: Given that, $n = 5000$, $p = 0.001$, so $\lambda = np = 5000 \times 0.001 = 5$. Thus the probability that the company will have to pay at least 2 claims during a given year is given by

$$\begin{aligned} P(x \geq 2) &= 1 - P(x < 2) = 1 - [P(x=0) + P(x=1)] \\ &= 1 - [e^{-\lambda} + \lambda e^{-\lambda}] = 1 - [e^{-5} + 5e^{-5}] = 1 - 6e^{-5} \\ &= 1 - 6 \times 0.0067 = 0.9598 \end{aligned}$$

Example 6.17: A manufacturer who produces medicine bottles, finds that 0.1 per cent 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 Poisson distribution, find how many boxes will contain:

(i) no defectives

(ii) at least two defectives [Delhi Univ., MBA, 1996, 2001]

Solution: Given that, $p = 1$ per cent $= 0.001$, $n = 500$, $\lambda = np = 500 \times 0.001 = 0.5$

$$(i) P[x=0] = e^{-\lambda} = e^{-0.5} = 0.6065$$

Therefore, the required number of boxes are : $0.6065 \times 100 = 61$ (approx.)

$$\begin{aligned} (ii) P(x > 2) &= 1 - P(x \leq 1) = 1 - [P(x=0) + P(x=1)] \\ &= 1 - [e^{-\lambda} + \lambda e^{-\lambda}] = 1 - [0.6065 + 0.5(0.6065)] \\ &= 1 - 0.6065 (1.5) = 1 - 0.90975 = 0.09025. \end{aligned}$$

Therefore, the required number of boxes are $100 \times 0.09025 = 10$ (approx.)

Example 6.18: The following table gives the number of days in a 50-day period during which automobile accidents occurred in a city :

No. of accidents	: 0	1	2	3	4
No. of days	: 21	18	7	3	1

Fit a Poisson distribution to the data.

[Sukhadia Univ.,MBA,1992; Kumaon Univ., MBA,2000]

Solution: Calculations for fitting of Poisson distribution are shown in the Table 6.3.

Table 6.5: Calculations for Poisson Distribution

Number of Accidents (x)	Number of Days (f)	f_x
0	21	0
1	18	18
2	7	14
3	3	09
4	1	04
	$n = 50$	$\Sigma f_x = 45$

Thus \bar{x} (or λ) $= \frac{\Sigma f x}{n} = \frac{45}{50} = 0.9$

and

$$P(x=0) = e^{-\lambda} = e^{-0.9} = 0.4066$$

$$P(x=1) = \lambda P(x=0) = 0.9(0.4066) = 0.3659$$

$$P(x=2) = \frac{\lambda}{2} P(x=1) = \frac{0.9}{2} (0.3659) = 0.1647$$

$$P(x=3) = \frac{\lambda}{3} P(x=2) = \frac{0.9}{3} (0.1647) = 0.0494$$

$$P(x=4) = \frac{\lambda}{4} P(x=3) = \frac{0.9}{4} (0.0494) = 0.0111$$

In order to fit a Poisson distribution, we shall multiply each of these values by $N = 50$ (total frequencies). Hence the expected frequencies are:

$x :$	0	1	2	3	4
$f :$	0.4066×50 = 20.33	0.3659×50 = 18.30	0.1647×50 = 8.23	0.0494×50 = 2.47	0.0111×50 = 0.56

6.5.3 Negative Binomial Probability Distribution

All conditions of binomial distribution are also applicable to the negative binomial distribution except that it describes the number of trials likely to be required to obtain a fixed number of successes. For example, suppose a percentage p of individuals in the population are sampled until exactly r individuals with the certain characteristic are found. The number of individuals in excess of r that are observed or sampled has a negative binomial distribution.

The probability distribution function of the negative binomial distribution is obtained by considering an infinite series of Bernoulli trials with probability of success p of an event on an individual trial. If trials are repeated r times until an event of interest occurs, then the probability that at least m trials will be required to get the event r times (successes) is given by

$$\begin{aligned} P(m, r, p) &= \text{Probability that an event occurs } (r-1) \text{ times in the first } m-1 \text{ trials} \\ &\quad \times \text{Probability that the event of interest occurs in the } m\text{th trial} \\ &= {}^{m-1}C_{r-1} p^{r-1} q^{m-r} \times p = {}^{m-1}C_{r-1} p^r q^{m-r}, \quad m = r, r+1, \dots \end{aligned} \quad (6-5)$$

where $r \geq 1$ is a fixed integer.

A random variable having a negative binomial distribution is also referred to as a discrete waiting time random variable. In terms of number of failures, it represents how long one waits for the r th success.

The mean and variance of this distribution are given by

$$\text{Mean, } \mu = \frac{r}{p}, \quad \text{Variance, } \sigma^2 = \frac{rq}{p^2}$$

Example 6.19: A market research agency that conducts interviews by telephone has found from past experience that there is a 0.40 probability that a call made between 2.30 PM and 5.30 PM will be answered. Assuming a Bernoullian process,

- (a) calculate the probability that an interviewer's 10th answer comes on his 20th call and that he will receive the first answer on his 3rd call,
- (b) what is the expected number of calls necessary to obtain seven answers.

Solution: Let answer to a call be considered 'success'. Then $p = 0.40$

$$\begin{aligned} (a) \quad P(10\text{th answer comes on 20th call}) \\ &= {}^{m-1}C_{r-1} p^r q^{m-r}, \quad m = r, r+1, \dots \\ &= {}^{19}C_9 (0.4)^{10} (0.6)^{10}; \quad m = 20 \text{ and } r = 10 \\ &= 0.058 \end{aligned}$$

$$P(\text{First answer on 3rd call}) = {}^2C_0 (0.4)^1 (0.6)^2 = 0.144$$

- (b) Expected number of calls for 7 answers, $\mu = r/p = 7/0.4 = 17.5 \approx 18$ calls

6.5.4 Multinomial Probability Distribution

The binomial distribution discussed earlier is associated with a sequence of n independent repeated Bernoulli trials, each resulting in only two outcomes, and one of the possible outcomes is called success, while multinomial distribution is associated with independent repeated trials that generalize from Bernoulli trials each resulting in two to k outcomes.

Suppose a single trial of an experiment results in only one of the k possible outcomes O_1, O_2, \dots, O_k with respective probabilities p_1, p_2, \dots, p_k ; and the experiment is repeated n times independently. The probability that out of these n trials outcome O_1 occurs x_1 times, O_2 occurs x_2 times and so on, is given by the following discrete density function:

$$P(x_1, x_2, \dots, x_k) = \frac{n!}{x_1! x_2! \dots x_k!} [p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}] \quad (6-6)$$

where $x_1 + x_2 + \dots + x_k = n$.

Example 6.20: In a factory producing certain items, 30 per cent of the items produced have no defect, 40 per cent have one defect, and 30 per cent have two defects. A random sample of 8 items is taken from a day's output. Find the probability that it will contain 2 items with no defect, 3 items with one defect, and 3 items with two defects.

Solution: We know from the data that $n = 8, p_1 = 0.30, p_2 = 0.40$ and $p_3 = 0.30; x_1 = 2, x_2 = 3, x_3 = 3$. Thus the required probability is given by

$$\begin{aligned} P(x_1=2, x_2=3, x_3=3) &= \frac{8!}{2! 3! 3!} [(0.30)^2 (0.40)^2 (0.30)^3] \\ &= 0.0871 \end{aligned}$$

6.5.5 Hypergeometric Probability Distribution

For a binomial distribution to be applied, the probability of a success or failure must remain the same for each trial. This is possible only when the number of elements in the population are large relative to the number in the sample, where probability of getting a success on a single trial is equal to the proportion p of successes in the population. However, if the number of element in the population are small in relation to the sample size, i.e. $n/N \geq 0.5$, the probability of a success in a given trial is dependent upon the outcomes of preceding trials. Then the number r of successes follows hypergeometric probability distribution. Thus hypergeometric probability distribution is similar to binomial distribution where probability of success may be different from trial to trial. When sampling is done *without replacement* from a finite population, the Bernoulli process does not apply because there is a systematic change in the probability of success in the reduced size of population.

Let N be the size of population and out of N , m be the total number of elements having a certain characteristic (called success) and the remaining $N - m$ do not have it, such that $p + q = 1$. Suppose a sample of size n is drawn at random without replacement. Then in a random sample of size n , the probability of exactly r successes when values of r depend on N, p and n is given by

$$P(x = r) = \frac{{m \choose r}^{N-m} {N-r \choose N}}{N \choose n}; \quad r = 0, 1, 2, \dots, n; \quad \text{and } 0 \leq r \leq m$$

This probability mass function is called *hypergeometric probability distribution*.

The mean and variance of a hypergeometric distribution are

$$\text{Mean} = n \left(\frac{m}{N} \right) \text{ and Variance} = n \left(\frac{m}{N} \right) \left(\frac{N-m}{N} \right) \left(\frac{N-n}{N-1} \right)$$

Example 6.21: Suppose the HRD manager randomly selects 3 individuals from a group of 10 employees for a special assignment. Assuming that 4 of the employees were assigned to a similar assignment previously, determine the probability that exactly two of the three employees have had previous experience.

Solution: We know from the data that $N = 10, n = 3, r = 2, m = 4$, and $N - m = 6$. Thus the required probability is given by

$$\begin{aligned} P(x = r | N, m, N-m) &= \frac{\frac{m}{N} C_r^{N-m} C_{n-r}}{N C_n} = \frac{\frac{4}{10} C_2^{10-4} C_{3-2}}{10 C_3} \\ &= \frac{\frac{4}{10} C_2^6 C_1}{10 C_3} = \frac{\left(\frac{4!}{2!2!}\right)\left(\frac{6!}{1!5!}\right)}{\left(\frac{10!}{3!7!}\right)} = \frac{36}{120} = 0.30 \end{aligned}$$

Example 6.22: Suppose a particular industrial product is shipped in lots of 20. To determine whether an item is defective a sample of 5 items from each lot is drawn. A lot is rejected if more than one defective item is observed. (If the lot is rejected, each item in the lot is then tested). If a lot contains four defectives, what is the probability that it will be accepted?

Solution: Let r be the number of defectives in the sample size $n = 5$. Given that, $N = 20$, $m = 4$, and $N - m = 16$. Then

$$\begin{aligned} P(\text{accept the lot}) &= P(x \leq 1) = P(x = 0) + P(x = 1) \\ &= \frac{4C_0 \times 16C_5}{20C_5} + \frac{4C_1 \times 16C_4}{20C_5} = \frac{\frac{4!}{0!4!} \times \frac{16!}{5!11!}}{\frac{20!}{5!15!}} + \frac{\frac{4!}{1!3!} \times \frac{16!}{4!12!}}{\frac{20!}{5!15!}} \\ &= \frac{91}{323} + \frac{455}{969} = 0.2817 + 0.4696 = 0.7513 \end{aligned}$$

Conceptual Questions 6C

13. If x has a Poisson distribution with parameter λ , then show that $E(x)$ and $V(x) = \lambda$. Further, show that the Poisson distribution is a limiting form of the binomial distribution.
14. What is Poisson distribution? Point out its role in business decision-making. Under what conditions will it tend to become a binomial distribution?
[Kumaon Univ., MBA, 1998]
15. When can Poisson distribution be a reasonable approximation of the binomial?
[Delhi Univ., MCom, 1999]

16. Discuss the distinctive features of Poisson distribution. When does a binomial distribution tend to become a Poisson distribution?
17. Under what conditions is the Poisson probability distribution appropriate? How are its mean and variance calculated?
18. What is negative binomial distribution? Distinguish the relationship between the binomial and negative binomial distributions.
19. What is hypergeometric distribution? Explain its properties.

Self-Practice Problems 6C

- 6.24 The following table shows the number of customers returning the products in a marketing territory. The data is for 100 stores:

No. of returns :	0	1	2	3	4	5	6
No. of stores :	4	14	23	23	18	9	9

Fit a Poisson distribution. [Lucknow Univ., MBA, 1997]

- 6.25 In a certain factory manufacturing razor blades, there is a small chance of 1/150 for any blade to be defective. The blades are placed in packets, each containing 10 blades. Using the Poisson distribution, calculate the approximate number of packets containing not more than 2 defective blades in a consignment of 10,000 packets.

- 6.26 In a town 10 accidents took place in a span of 50 days. Assuming that the number of accidents per day follows the Poisson distribution, find the probability that there will be three or more accidents in a day.
[Coimbatore Univ., MBA, 1997]

- 6.27 The distribution of typing mistakes committed by a typist is given below. Assuming a Poisson distribution, find out the expected frequencies:

No. of mistakes	per page	:	0	1	2	3	4	5
	No. of pages :		142	156	69	27	5	1

[Rohilkhand Univ., MBA, 1998]

- 6.28 Find the probability that at most 5 defective bolts will be found in a box of 200 bolts if it is known that 2 per cent of such bolts are expected to be defective [you may take the distribution to be Poisson; $e^{-4} = 0.0183$].

- 6.29 On an average, one in 400 items is defective. If the items are packed in boxes of 100, what is the probability that any given box of items will contain: (i) no defectives; (ii) less than two defectives; (iii) one or more defectives; and (iv) more than three defectives [Delhi Univ., MBA, 2000]

6.30 It is given that 30 per cent of electric bulbs manufactured by a company are defective. Find the probability that a sample of 100 bulbs will contain (i) no defective, and (ii) exactly one defective.

6.31 One-fifth per cent of the blades produced by a blade manufacturing factory turn out to be defective. The blades are supplied in packets of 10. Use Poisson distribution to calculate the approximate number of packets containing no defective, one defective, and two defective blades respectively in a consignment of 1,00,000 packets. [Delhi Univ., MBA, 1999, 2003]

6.32 A factory produces blades in packets of 10. The probability of a blade to be defective is 0.2 per cent. Find the number of packets having two defective blades in a consignment of 10,000 packets.

6.33 When a first proof of 200 pages of an encyclopaedia of 5,000 pages was read, the distribution of printing mistakes was found to be as shown in the first and second columns of the table below. Fit a Poisson distribution to the frequency distribution of printing mistakes. Estimate the total cost of correcting the whole encyclopaedia by using the information given in the first and third columns of the table below:

Misprints on a Page	Frequency	Cost of Detection and Correction Per Page (Rs)
0	113	1.00
1	62	2.50
2	20	1.50
3	3	3.00
4	1	3.50
5	1	4.00

[MD Univ., MBA, 1998]

6.34 In a certain factory manufacturing razor blades, there is small chance 1/50 for any blade to be defective. The blades are placed in packets, each containing 10 blades. Using an appropriate probability distribution, calculate the approximate number of packets containing not more than 2 defective blades in a consignment of 10,000 packets.

6.35 Suppose that a manufactured product has 2 defects per unit of product inspected. Using Poisson distribution, calculate the probabilities of finding a product without any defect, 3 defects, and 4 defects. (Given $e^{-2} = 0.135$)

[Madurai Univ., MCom, 1999]

6.36 A distributor received a shipment of 12 TV sets. Shortly after this shipment was received, the manufacturer informed that he had inadvertently shipped 3 defective sets. The distributor decided to test 4 sets randomly selected out of 12 sets received.

(a) What is the probability that neither of the 4 sets tested was defective?

(b) What is the mean and variance of defective sets.

6.37 Suppose a population contains 10 elements, 6 of which are defective. A sample of 3 elements is selected. What is the probability that exactly 2 are defective?

6.38 A transport company has a fleet of 15 trucks, used mainly to deliver fruits to wholesale market. Suppose 6 of the 15 trucks have brake problems. Five trucks were selected at random to be tested. What is the probability that 2 of those tested trucks have defective brakes?

6.39 A company has five applicants for two positions: two women and three men. Suppose that the five applicants are equally qualified and that no preference is given for choosing either gender. If r equal the number of women chosen to fill the two positions, then what is the probability distribution of r . Also, determine the mean and variance of this distribution.

Hints and Answers

6.24 Fitting of Poisson distribution

$$\begin{array}{ccccccc} x & : & 0 & 1 & 2 & 3 & 4 \\ f & : & 4 & 14 & 23 & 23 & 18 \\ fx & : & 0 & 14 & 46 & 69 & 72 \end{array} \quad \begin{array}{c} 5 \\ 9 \\ 9 = 100 \end{array}$$

$$\therefore \lambda = 300/100 = 3. \text{ Then}$$

$$\begin{aligned} NP(x=0) &= 100 e^{-\lambda} = 100(2.7183)^{-3} = 5; \\ P(x=1) &= \lambda NP(0) = 15 \end{aligned}$$

$$P(x=2) = NP(x=1) \frac{\lambda}{2} = 22.5;$$

$$P(x=3) = NP(x=2) \frac{\lambda}{3} = 22.5$$

$$P(x=4) = NP(x=3) \frac{\lambda}{4} = 16.9;$$

$$P(x=5) = NP(x=4) \frac{\lambda}{5} = 10.1$$

$$P(x=6) = NP(x=5) \frac{\lambda}{6} = 5.1$$

6.25 Given that $N = 10,000$, $p = 1/50$, $n = 10$, $\lambda = np = 10 \times (1/50) = 0.2$

$$P(x=0) = e^{-\lambda} = e^{-0.2} = 0.8187 \text{ (from the table)}$$

$$NP(x=0) = 0.8187 \times 10,000 = 8187$$

$$NP(x=1) = NP(x=0) \times \lambda = 8187 \times 0.2 = 1637.4$$

$$\begin{aligned} NP(x=2) &= NP(x=1) \times \lambda/2 = 1637.4 \times (0.2/2) \\ &= 163.74 \end{aligned}$$

The approximate number of packets containing not more than 2 defective blades in a consignment of 10,000 packets is: $10,000 - (8187 + 163.74 + 163.74) = (10,000 - 9988.14) = 11.86$ or 12.

6.26 The average number of accidents per day = $10/50 = 0.2$

$$\begin{aligned} P(x \geq 3 \text{ accidents}) &= 1 - P(2 \text{ or less accidents}) \\ &= 1 - [P(0) + P(1) + P(2)] \end{aligned}$$

$$\begin{aligned}
 &= -\left[e^{-2} + 2e^{-2} + \frac{e^{-2} \times 0.2 \times 0.2}{2} \right] \\
 &= 1 - e^{-2} [1 + 0.2 + 0.02] = 1 - 0.8187 \times 1.22 \\
 &\quad \text{(From table of } e^{-\lambda}) \\
 &= 1 - 0.998 = 0.002
 \end{aligned}$$

6.27 $\lambda = \sum fx / N = 400/400 = 1,$
 $P(x = 0) = e^{-\lambda} = 0.3679$

$$\begin{aligned}
 NP(x = 0) &= 147.16 \\
 NP(x = 1) &= NP(x = 0)\lambda = 147.16 \\
 NP(x = 2) &= NP(x = 1) \frac{\lambda}{2} = 73.58 \\
 NP(x = 3) &= NP(x = 2) \frac{\lambda}{3} = 24.53 \\
 NP(x = 4) &= NP(x = 3) \frac{\lambda}{4} = 6.13 \\
 NP(x = 5) &= NP(x = 4) \frac{\lambda}{5} = 1.23
 \end{aligned}$$

Expected frequencies as per the distribution are:

No. of mistakes per page	: 0 1 2 3 4 5
No. of pages	: 147 147 74 25 6 1

6.28 $p(\text{defective bolt}) = 2\% = 0.02.$ Given $n = 200,$ so $\lambda = np = 200 \times 0.02 = 4$

$$P(0) = \frac{e^{-\lambda} \lambda^0}{0!} = e^{-4} = 0.0183$$

$$\begin{aligned}
 P(x \leq 5) &= P(x = 0) + P(x = 1) + P(x = 2) + \\
 &\quad P(x = 3) + P(x = 4) + P(x = 5) \\
 &= e^{-4} \left(1 + 4 + \frac{4^2}{2!} + \frac{4^3}{3!} + \frac{4^4}{4!} + \frac{4^5}{5!} \right) \\
 &= 0.0183 \times (643/15) = 0.7844.
 \end{aligned}$$

6.30 $\lambda = np = 100 \times 0.30 = 3$

$$P(x = 0) = e^{-\lambda} = e^{-3} = 0.05;$$

$$P(x = 1) = \lambda P(x = 0) = 3 \times 0.03 = 0.15$$

6.31 Given $n = 10, p = 1/500, \lambda = np = 10/500 = 0.02$

(i) $P(x = 0) = e^{-\lambda} = e^{-0.02} = 0.9802$

$$\begin{aligned}
 NP(x = 0) &= 1,00,000 \times 0.9802 \\
 &= 98020 \text{ packets}
 \end{aligned}$$

(ii) $P(x = 1) = \lambda P(x = 0)$

$$\begin{aligned}
 &= \lambda e^{-\lambda} = 0.02 \times 0.9802 \\
 &= 0.019604
 \end{aligned}$$

$$\begin{aligned}
 NP(x = 1) &= 1,00,000 \times 0.019604 \\
 &= 1960 \text{ packets}
 \end{aligned}$$

$$\begin{aligned}
 \text{(iii)} \quad P(x = 2) &= \frac{\lambda^2}{2} P(x = 0) = \frac{(0.02)^2}{2} \times 0.9802 \\
 &= 0.00019604
 \end{aligned}$$

$$\begin{aligned}
 NP(x = 2) &= 1,00,000 \times 0.00019604 \\
 &= 19.60 \approx 20 \text{ packets}
 \end{aligned}$$

6.32 Given $n = 10, p = 0.002, \lambda = np = 10 \times 0.002 = 0.02.$

$$\begin{aligned}
 P(x = 2) &= \frac{e^{-\lambda} \lambda^2}{2!} = \frac{e^{-0.02}(0.02)^2}{2!} \\
 &= 0.000196
 \end{aligned}$$

The required number of packets having two defective blades each in a consignment of 10,000 packets
 $= 10,000 \times 0.000196 \approx 2.$

6.33 No. of mis-

Prints (x)	:	0	1	2	3	4	5
Frequency	:						
(f)	:	113	62	20	3	1	1
fx	:	0	62	40	09	04	05

$$\therefore \bar{x} = \sum fx/N = 120/200 = 0.6 (= \lambda)$$

For fitting of Poisson distribution, calculating

$$\begin{aligned}
 NP(x = 0) &= 200 \left[\frac{e^{-\lambda} \lambda^0}{0!} \right] = 200 e^{-0.6} \\
 &= 200 (0.5488) = 109.76 \\
 NP(x = 1) &= 200 P_0 \times \lambda = 200 \times 109.76 \times 0.6 \\
 &= 65.856
 \end{aligned}$$

$$\begin{aligned}
 NP(x = 2) &= 200 P_1 \times \frac{\lambda}{2} = 65.856 \times \frac{0.6}{2} \\
 &= 19.756
 \end{aligned}$$

$$\begin{aligned}
 NP(x = 3) &= 200 P_2 \times \frac{\lambda}{3} = 19.756 \times \frac{0.6}{3} \\
 &= 3.951
 \end{aligned}$$

$$\begin{aligned}
 NP(x = 4) &= 200 P_3 \times \frac{\lambda}{4} = 3.951 \times \frac{0.6}{4} \\
 &= 0.5927
 \end{aligned}$$

$$\begin{aligned}
 NP(x = 5) &= 200 P_4 \times \frac{\lambda}{5} = 0.5927 \times \frac{0.6}{5} \\
 &= 0.0711
 \end{aligned}$$

The total cost of correcting the first proof of the whole encyclopaedia will be

No. of Misprints/Page	Rate/Page (x)	No. of Pages (f)	Total Cost of Correcting Proof (fx)
0	1.00	109.760	109.760
1	1.50	65.856	98.784
2	2.50	19.756	49.390
3	3.00	3.951	11.853
4	3.50	0.592	2.074
5	4.00	0.071	0.284
			Rs 272.145

6.34 Given $N = 10,000, p = 1/50$ and $n = 10.$

Thus $\lambda = np = 0.20$ and

$$\begin{aligned}
 NP(x = 0) &= 10,000 e^{-\lambda} = 10,000 e^{-0.20} = 8187; \\
 P(x = 1) &= NP(x = 0) \times \lambda = 8187 \times 0.2 = 1637.4.
 \end{aligned}$$

$$\begin{aligned}
 NP(x = 2) &= NP(x = 1) \times \frac{\lambda}{2} = 1637.4 \times \frac{0.2}{2} \\
 &= 163.74
 \end{aligned}$$

The approximate number of packets containing not more than 2 defective blades in a consignment of 10,000 packets will be: $10,000 - (8187 + 1637.40 + 163.74) = 12$ approx.

6.35 Given average number of defects, $\lambda = 2$.

$$P(x=0) = e^{-\lambda} = e^{-2} = 0.135;$$

$$P(x=1) = P(x=0) \times \lambda = 0.135 \times 2 = 0.27$$

$$P(x=2) = P(x=1) \times \frac{\lambda}{2} = 0.27 \times \frac{2}{2} = 0.27$$

$$P(x=3) = P(x=2) \times \frac{\lambda}{3} = 0.27 \times \frac{2}{3} = 0.18$$

$$P(x=4) = P(x=3) \times \frac{\lambda}{4} = 0.18 \times \frac{2}{4} = 0.09$$

6.36 Given $N = 12$, $n = 4$, $m = 3$ and $N - m = 9$.

$$(a) P(x=0) = \frac{^3C_0 \times ^9C_4}{^{12}C_4} = \frac{\frac{3!}{0!3!} \times \frac{9!}{4!5!}}{\frac{12!}{4!8!}} = \frac{1 \times 126}{495} = \frac{14}{55}$$

$$(b) \text{Mean}, \mu = n \left(\frac{m}{N} \right) = 4 \left(\frac{3}{12} \right) = 1$$

$$\begin{aligned} \text{Variance}, \sigma^2 &= n \left(\frac{m}{N} \right) \left(\frac{N-m}{N} \right) \left(\frac{N-n}{N-1} \right) \\ &= 4 \left(\frac{3}{12} \right) \left(\frac{9}{12} \right) \left(\frac{8}{11} \right) = 0.5455 \end{aligned}$$

$$\text{6.37 } P(x=2) = \frac{^6C_2 \times ^4C_1}{^{10}C_3} = \frac{15 \times 4}{120} = 0.50$$

$$\text{6.38 } P(x=2) = \frac{^9C_3 \times ^6C_2}{^{15}C_5} = \frac{84 \times 15}{3003} = 0.4196$$

6.39 Given $N = 5$, $n = 2$, $m = 2$, $N - m = 3$

$$P(x=r) = \frac{^mC_r \times ^{N-m}C_{n-r}}{^NC_n} = \frac{^2C_r \times ^3C_{2-r}}{^5C_2}; r = 0, 1, 2$$

$$\text{Mean, } \mu = 2 \left(\frac{2}{5} \right) = 0.8;$$

$$\text{Variance} = 2 \left(\frac{2}{5} \right) \left(\frac{3}{5} \right) \left(\frac{3}{4} \right) = 0.6$$

6.6 CONTINUOUS PROBABILITY DISTRIBUTIONS

If a random variable is discrete, then it is possible to assign a specific probability to each of its value and get the probability distribution for it. The sum of all the probabilities associated with the different values of the random variable is 1. However, not all experiments result in random variables that are discrete. Continuous random variables such as height, time, weight, monetary values, length of life of a particular product, etc. can take large number of observable values corresponding to points on a line interval much like the infinite number of gains of sand on a beach. The sum of probability to each of these infinitely large values is no longer sum to 1.

Unlike discrete random variables, continuous random variables do not have probability distribution functions specifying the exact probabilities of their specified values. Instead, probability distribution is created by distributing one unit of probability along the real line, much like distributing a handful of sand along a line. The probability of measurements (e.g. gains of sand) piles up in certain places resulting into a probability distribution called *probability density function*. Such distribution is used to find probabilities that the random variable falls into a specified interval of values. The depth or density of the probability that varies with the random variable (x) may be described by a mathematical formula.

The probability density function for a continuous random variable x is a curve such that the area under the curve over an interval equals the probability that x falls into that interval, i.e. the probability that x is in that interval can be found by summing the probabilities in that interval. Certain characteristics of probability density function for the continuous random variable, x are follows:

- (i) The area under a continuous probability distribution is equal to 1.
- (ii) The probability $P(a \leq x \geq b)$ that random variable x value will fall into a particular interval from a to b is equal to the area under the density curve between the points (values) a and b .

Nature seems to follow a predictable pattern for many kinds of measurements. Most numerical values of a random variable are spread around the center, and greater the distance a numerical value has from the center, the fewer numerical values have that

Normal distribution: A continuous probability distribution in which the curve is bell-shaped having a single peak. The mean of the distribution lies at the center of the curve and the curve is symmetrical around a vertical line erected at the mean. The tails of the curve extend indefinitely parallel to the horizontal axis.

specific value. A frequency distribution of values of random variable observed in nature which follows this pattern is approximately bell shaped. A special case of distribution of measurements is called a **normal curve (or distribution)**.

If a population of numerical values follows a normal curve and x is the randomly selected numerical value from the population, then x is said to be *normal random variable*, which has a normal probability distribution.

W.J. Youden, a well-known statistician expressed his views about normal distribution as follows:

THE
NORMAL
LAW OF ERROR
STANDS OUT IN
THE EXPERIENCE OF
MANKIND AS ONE OF THE
BROADEST GENERALISATION OF
NATURAL PHILOSOPHY. IT SERVES AS THE
GUIDING INSTRUMENT RESEARCHES IN THE
PHYSICAL AND SOCIAL SCIENCES AND IN MEDICINE
AGRICULTURE AND ENGINEERING. IT IS AN INDISPENSABLE
TOOL FOR THE ANALYSIS AND THE INTERPRETATION OF
THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT

Artistically, it also look like a normal curve

The normal distribution also known as *Gaussian distribution* is due to the work of German mathematician Karl Friedrich Gauss during the early part of the 19th century. Normal distribution provides an adequate representation of a continuous phenomenon or process such as daily changes in the stock market index, frequency of arrivals of customers at a bank, frequency of telephone calls into a switch board, customer servicing times, and so on.

6.6.1 Normal Probability Distribution Function

The formula that generates normal probability distribution is as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{(-1/2)[(x-\mu)/\sigma]^2}, \quad -\infty < x < \infty \quad (6-6)$$

where π = constant 3.1416

e = constant 2.7183

μ = mean of the normal distribution

σ = standard of normal distribution

The $f(x)$ values represent the relative frequencies (height of the curve) within which values of random variable x occur. The graph of a normal probability distribution with mean μ and standard deviation σ is shown in Fig. 6.7. The distribution is symmetric about its mean μ that locates at the centre.

Since the total area under the normal probability distribution is equal to 1, the symmetry implies that the area on either side of μ is 50 per cent or 0.5. The *shape* of the distribution is determined by μ and σ values.

In symbols, if a random variable x follows normal probability distribution with mean μ and standard deviation σ , then it is also expressed as: $x \sim N(\mu, \sigma)$.

Characteristics of the Normal Probability Distribution There is a family of normal distributions. Each normal distribution may have a different mean μ or standard deviation σ . A unique normal distribution may be defined by assigning specific values to the mean

μ and standard deviation σ in the normal probability density function (6-6). Large value of σ reduce the height of the curve and increase the spread; small values of σ increase the height of the curve and reduce the spread. Figure 6.6(a) shows three normal distributions with different values of the mean μ and a fixed standard deviation σ , while in Fig. 6.6(b) normal distributions are shown with different values of the standard deviation σ and a fixed mean μ .

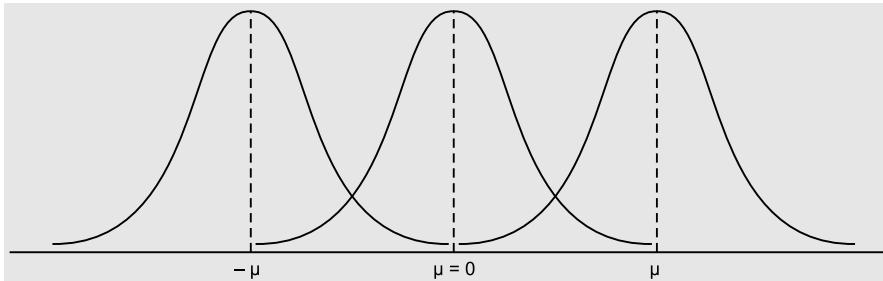


Figure 6.6 (a)
Normal Distributions with Different Mean Values But Fixed Standard Deviation

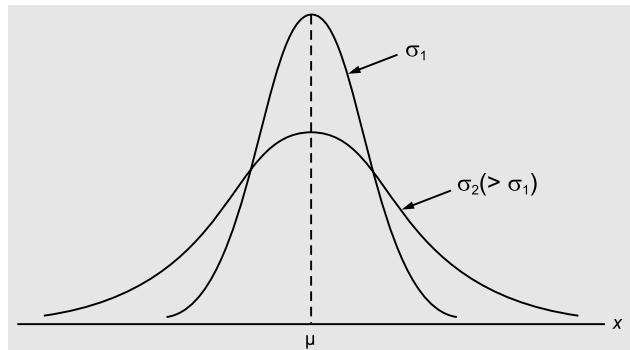


Figure 6.6 (b)
Normal Distributions with Fixed Mean and Variable Standard Deviation

From Fig. 6.6(a) and 6.6(b) the following characteristics of a normal distribution and its density function may be derived:

- (i) For every pair of values of μ and σ , the curve of normal probability density function is bell shaped and symmetric.
- (ii) The normal curve is symmetrical around a vertical line erected at the mean μ with respect to the area under it, that is, fifty per cent of the area of the curve lies on both sides of the mean and reflect the mirror image of the shape of the curve on both sides of the mean μ . This implies that the probability of any individual outcome above or below the mean will be same. Thus, for any normal random variable x ,

$$P(x \leq \mu) = P(x \geq \mu) = 0.50$$

- (iii) Since the normal curve is symmetric, the mean, median, and mode for the normal distribution are equal because the highest value of the probability density function occurs when value of a random variable, $x = \mu$.
- (iv) The two tails of the normal curve extend to infinity in both directions and theoretically never touch the horizontal axis.
- (v) The mean of the normal distribution may be negative, zero, or positive as shown in Fig. 6.6(a).
- (vi) The mean μ determines the *central location* of the normal distribution, while standard deviation σ determines its *spread*. The larger the value of the standard deviation σ , the wider and flatter is the normal curve, thus showing more variability in the data, as shown in Fig. 6.6(b). Thus standard deviation σ determines the range of values that any random variable is likely to assume.
- (vii) The area under the normal curve represents probabilities for the normal random variable, and therefore, the total area under the curve for the normal probability distribution is 1.

Standard Normal Probability Distribution: To deal with problems where the normal probability distribution is applicable more simply, it is necessary that a random variable x is standardized by expressing its value as the number of standard deviations (σ) it lies to the left or right of its mean (μ). The *standardized normal random variable*, z (also called *z-statistic*, *z-score* or *normal variate*) is defined as:

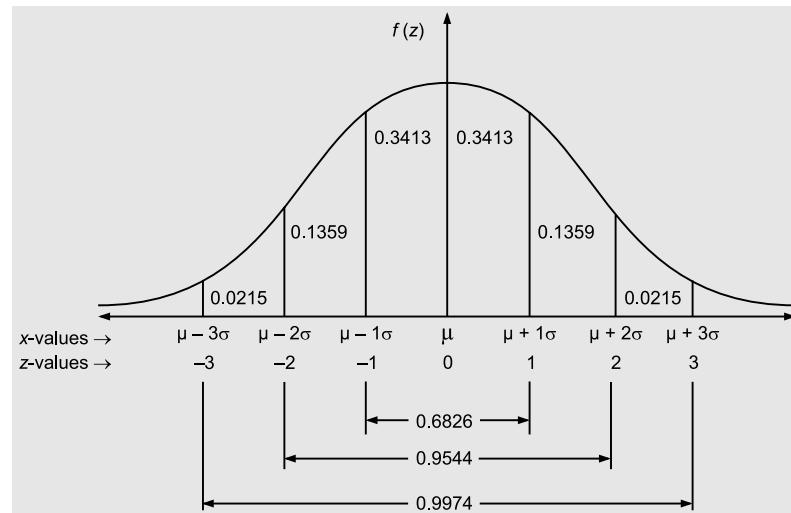
$$z = \frac{x - \mu}{\sigma} \quad (6-7)$$

or equivalently $x = \mu + z\sigma$

A z -score measures the number of standard deviations that a value of the random variable x fall from the mean. From formula (6.7) we may conclude that

- (i) When x is less than the mean (μ), the value of z is negative
- (ii) When x is more than the mean (μ), the value of z is positive
- (iii) When $x = \mu$, the value of $z = 0$.

Figure 6.7
Standard Normal Distribution

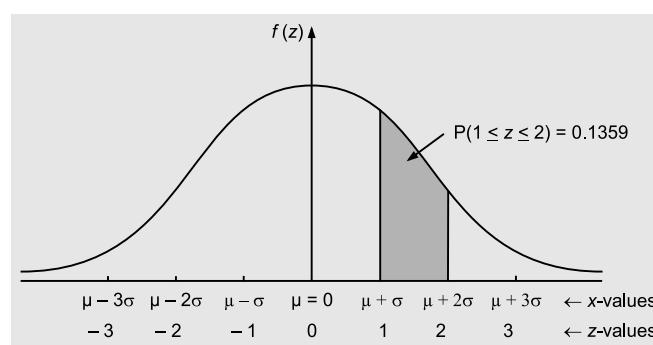


Any normal probability distribution with a set of μ and σ value with random variable can be converted into a distribution called **standard normal probability distribution** z , as shown in Fig. 6.7, with mean $\mu_z = 0$ and standard deviation $\sigma_z = 1$ with the help of the formula (7-7).

A z -value measures the distance between a particular value of random variable x and the mean (μ) in units of the standard deviation (σ). With the value of z obtained by using the formula (6.7), we can find the area or probability of a random variable under the normal curve by referring to the standard distribution in Appendix. For example $z = \pm 2$ implies that the value of x is 2 standard deviations above or below the mean (μ).

Area Under the Normal Curve Since the range of normal distribution is infinite in both the directions away from μ , the *pdf* function $f(x)$ is never equal to zero. As x moves away from μ , $f(x)$ approaches x -axis but never actually touches it.

Figure 6.8
Diagram for Finding $P(1 < z < 2)$



The area under the standard normal distribution between the mean $z = 0$ and a specified positive value of z , say z_0 is the probability $P(0 \leq z \leq z_0)$ and can be read off directly from standard normal (z) tables. For example, area between $1 \leq z \leq 2$ is the proportion of the area under the curve which lies between the vertical lines erected at two points along the x -axis. A portion of the table is shown in Table 6.6. For example, as shown in Fig. 6.8, if x is σ away from μ , that is, the distance between x and μ is one standard deviation or $(x - \mu)/\sigma = 1$, then 34.134 per cent of the distribution lies between x and μ . Similarly, if x is at

2σ away from μ , that is, $(x - \mu)/\sigma = 2$, then the area will include 47.725 per cent of the distribution, and so on, as shown in Table 6.6.

Table 6.6: Area Under the Normal Curve

$z = \frac{x - \mu}{\sigma}$	<i>Area Under Normal Curve Between x and μ</i>
1.0	0.34134
2.0	0.47725
3.0	0.49875
4.0	0.49997

Since the normal distribution is symmetrical, Table 6.6 indicates that about 68.26 per cent of the normal distribution lies within the range $\mu - \sigma$ to $\mu + \sigma$. The other relationships derived from Table 6.6 are shown in Table 6.7 and in Fig. 6.7.

Table 6.7: Percentage of the Area of the Normal Distribution Lying within the Given Range

<i>Number of Standard Deviations from Mean</i>	<i>Approximate Percentage of Area under Normal Curve</i>
$x \pm 2\sigma$	68.26
$x \pm 2\sigma$	95.45
$x \pm 3\sigma$	99.75

The standard normal distribution is a symmetrical distribution and therefore

$$P(0 \leq z \leq a) = P(-a \leq z \leq 0) \text{ for any value } a.$$

$$\begin{aligned} \text{For example, } P(1 \leq z \leq 2) &= P(z \leq 2) - P(z \leq 1) \\ &= 0.9772 - 0.8413 = 0.1359 \end{aligned}$$

The value of $P(1 \leq z \leq 2)$ is shown in Fig. 6.8.

6.6.2 Approximation of Binomial and Poisson Distributions to Normal Distribution

The binomial distribution approaches a normal distribution with standardized variable, that is,

$$\text{where } z = \frac{x - np}{\sqrt{npq}} \sim N(0, 1)$$

However this approximation works well when both $np \geq 10$ and $npq \geq 10$

Similarly, Poisson distribution also approaches a normal distribution with standardized variable, that is,

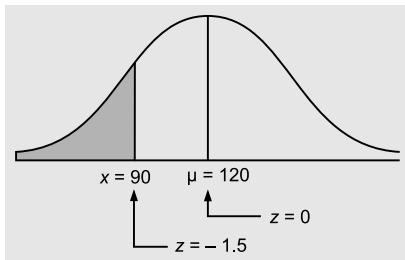
$$z = \frac{x - \lambda}{\sqrt{\lambda}} \sim N(0, 1)$$

Example 6.23: 1000 light bulbs with a mean life of 120 days are installed in a new factory and their length of life is normally distributed with standard deviation of 20 days.

- (a) How many bulbs will expire in less than 90 days?
- (b) If it is decided to replace all the bulbs together, what interval should be allowed between replacements if not more than 10% should expire before replacement?

Solution: (a) Given, $\mu = 120$, $\sigma = 20$, and $x = 90$. Then

$$z = \frac{x - \mu}{\sigma} = \frac{90 - 120}{20} = -1.5$$



The area under the normal curve between $z = 0$ and $z = -1.5$ is 0.4332. Therefore area to the left of -1.5 is $0.5 - 0.4332 = 0.0668$. Thus the expected number of bulbs to expire in less than 90 days will be $0.0668 \times 1000 = 67$ (approx.).

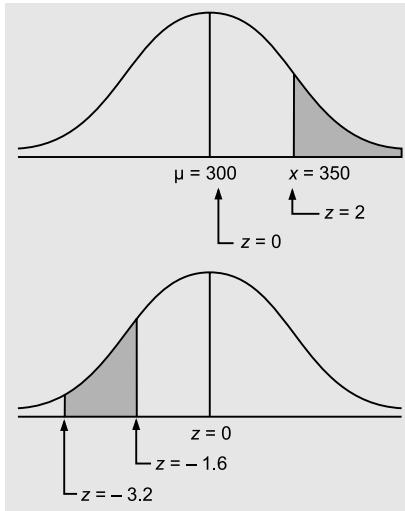
(b) The value of z corresponding to an area 0.4 ($0.5 - 0.10$). Under the normal curve is 1.28. Therefore

$$z = \frac{x - \mu}{\sigma} \text{ or } -1.28 = \frac{x - 120}{20} \text{ or } x = 120 - 20(-1.28) = 94$$

Hence, the bulbs will have to be replaced after 94 days.

Example 6.24: The lifetimes of certain kinds of electronic devices have a mean of 300 hours and standard deviation of 25 hours. Assuming that the distribution of these lifetimes, which are measured to the nearest hour, can be approximated closely with a normal curve

- (a) Find the probability that any one of these electronic devices will have a lifetime of more than 350 hours.
- (b) What percentage will have lifetimes of 300 hours or less?
- (c) What percentage will have lifetimes from 220 or 260 hours?



Solution: (a) Given, $\mu = 300$, $\sigma = 25$, and $x = 350$. Then

$$z = \frac{x - \mu}{\sigma} = \frac{350 - 300}{25} = 2$$

The area under the normal curve between $z = 0$ and $z = 2$ is 0.9772. Thus the required probability is, $1 - 0.9772 = 0.0228$.

$$(b) z = \frac{x - \mu}{\sigma} = \frac{300 - 300}{25} = 0$$

Therefore, the required percentage is, $0.5000 \times 100 = 50\%$.

(c) Given, $x_1 = 220$, $x_2 = 260$, $\mu = 300$ and $\sigma = 25$. Thus

$$z_1 = \frac{220 - 300}{25} = -3.2 \text{ and } z_2 = \frac{260 - 300}{25} = -1.6$$

From the normal table, we have

$$P(z = -1.6) = 0.4452 \text{ and } P(z = -3.2) = 0.4903$$

Thus the required probability is

$$P(z = -3.2) - P(z = -1.6) = 0.4903 - 0.4452 = 0.0541$$

Hence the required percentage = $0.0541 \times 100 = 5.41$ per cent.

Example 6.25: In a certain examination, the percentage of passes and distinctions were 46 and 9 respectively. Estimate the average marks obtained by the candidates, the minimum pass and distinction marks being 40 and 75 respectively (assume the distribution of marks to be normal).

Also determine what would have been the minimum qualifying marks for admission to a re-examination of the failed candidates, had it been desired that the best 25 per cent of them should be given another opportunity of being examined.

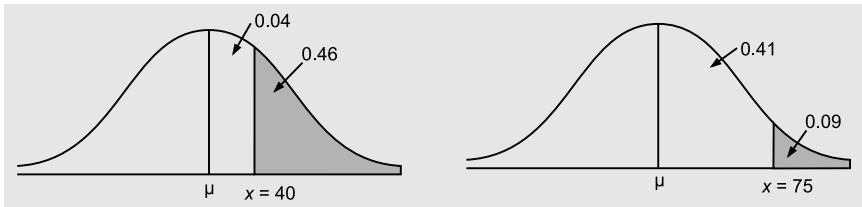
Solution: (a) Let μ be the mean and σ be the standard deviation of the normal distribution. The area to the right of the ordinate at $x = 40$ is 0.46 and hence the area between the mean and the ordinate at $x = 40$ is 0.04.

Now from the normal table, corresponding to 0.04, the standard normal variate, $z = 0.1$. Therefore, we have

$$\frac{40 - \mu}{\sigma} = 0.1 \text{ or } 40 - \mu = 0.1\sigma$$

$$\text{Similarly, } \frac{75 - \mu}{\sigma} = 1.34 \text{ or } 75 - \mu = 1.34\sigma$$

Solving these equations, we get $\sigma = 28.23$ and $\mu = 37.18$ or 37.



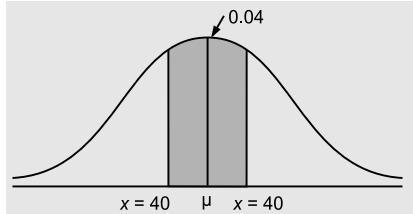
(b) Let us assume that x_1 is the minimum qualifying marks for re-examination of the failed candidates.

The area to the right of $x = 40$ is 46 per cent. Thus the percentage of students failing = 54 and this is the area to the left of 40. We want that the best 25 per cent of these failed candidates should be given a chance to reappear. Suppose this area is equal to the shaded area in the diagram. This area is, 25 per cent of 54 = 13.5 per cent = 0.1350.

The area between mean and ordinate at $x_1 = - (0.1350 - 0.04) = - 0.0950$ (negative sign is included because the area lies to the left of the mean ordinates).

Corresponding to this area, the standard normal variate $z = - 0.0378$. Thus, we write

$$\begin{aligned}\frac{x_1 - \mu}{\sigma} &= - 0.0378 \\x_1 &= \mu - 0.0378 \sigma \\&= 37.2 - (0.0378 \times 28.23) \\&= 37.2 - 1.067 = 36.133 \text{ or } 36 \text{ (approx.)}\end{aligned}$$



Example 6.26: In a normal distribution 31 per cent of the items are under 45 and 8 per cent are over 64. Find the mean and standard deviation of the distribution. [Delhi Univ., MBA, 1999]

Solution: Since 31 per cent of the items are under 45, therefore the left of the ordinate at $x = 45$ is 0.31, and obviously the area to the right of the ordinate up to the mean is $(0.5 - 0.31) = 0.19$. The value of z corresponding to this area is 0.5. Hence

$$z = \frac{45 - \mu}{\sigma} = - 0.5 \text{ or } -\mu + 0.5\sigma = - 45$$

As 8 per cent of the items are above 64, therefore area to the right of the ordinate at 64 is 0.08. Area to the left of the ordinate at $x = 64$ up to mean ordinate is $(0.5 - 0.08) = 0.42$ and the value of z corresponding to this area is 1.4. Hence

$$z = \frac{64 - \mu}{\sigma} = 1.4 \text{ or } -\mu - 1.4\sigma = - 64$$

From these two equations, we get $1.9\sigma = 19$ or $\sigma = 10$. Putting $\sigma = 10$ in the first equation, we get $\mu - 0.5 \times 10 = 45$ or $\mu = 50$.

Thus, mean of the distribution is 50 and standard deviation 10.

Example 6.27: The income of a group of 10,000 persons was found to be normally distributed with mean Rs 1750 p.m. and standard deviation Rs 50. Show that of this group 95% had income exceeding Rs 1668 and only 5 per cent had income exceeding Rs 1832. What was the lowest income among the richest 100? [Delhi Univ., MBA, 1997]

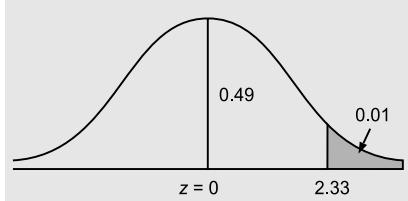
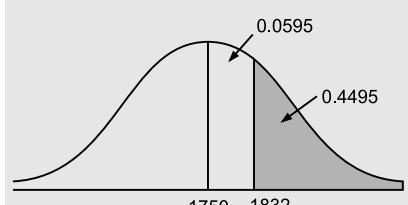
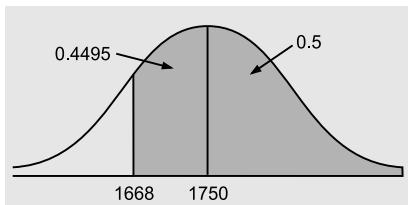
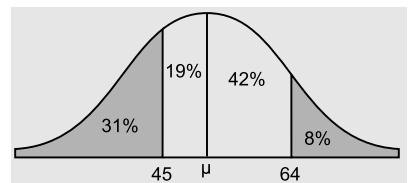
Solution: (a) Given that, $x = 1668$, $\mu = 1750$ and $\sigma = 50$. Therefore, the standard normal variate corresponding to $x = 1668$, is

$$z = \frac{x - \mu}{\sigma} = \frac{1668 - 1750}{50} = - 1.64$$

The area to the right of the ordinate at $z = - 1.64$ (or $x = 1668$) is $(0.4495 + 0.5000) = 0.9495$ (because $z = - 1.64$ to its right covers 95 per cent area).

The expected number of persons getting above Rs 1668 are $10,000 \times 0.9495 = 9495$. This is about 95 per cent of the total of 10,000 persons.

(ii) The standard normal variate corresponding to $x = 1832$ is



$$z = \frac{1832 - 1750}{50} = 1.64$$

The area to the right of ordinate at $z = 1.64$ is: $0.5000 - 0.4495 = 0.0505$

The expected number of persons getting above Rs 1832 is: $10,000 \times 0.0505 = 505$. This is about 5 per cent of the total of 10,000 persons. Thus probability of getting richest 100 out of 10,000 is $100/10,000 = 0.01$.

The standard normal variate having 0.01 area to its right is, $z = 2.33$. Hence

$$2.33 = \frac{x - 1750}{50}$$

$$x = 2.33 \times 50 + 1750 = \text{Rs } 1866 \text{ approx.}$$

This implies that the lowest among the richest 100 is getting Rs 1866 per month.

Example 6.28: A wholesale distributor of fertilizer products finds that the annual demand for one type of fertilizer is normally distributed with a mean of 120 tonnes and standard deviation of 16 tonnes. If he orders only once a year, what quantity should be ordered to ensure that there is only a 5 per cent chance of running short?

[Delhi Univ., MBA, 1998, 2000]

Solution: Let x be the annual demand (in tonnes) for one type of fertilizer. Therefore

$$z = \frac{x - 120}{16}$$

The desired area of 5 per cent is shown in the figure. Since the area between the mean and the given value of x is 0.45, therefore from the normal table this area of 0.45 corresponds to $z = 1.64$.

Substituting this value of $z = 1.64$ in standard normal variate, we get

$$1.64 = \frac{x - 120}{16}$$

$$\text{or } x = 120 + (1.64)(16) = 146.24 \text{ tonnes.}$$

If it is necessary to order in whole units, then the wholesale distributor should order 147 tonnes.

Example 6.29: Assume that the test scores from a college admissions test are normally distributed with a mean of 450 and a standard deviation of 100.

- What percentage of people taking the test score are between 400 and 500?
- Suppose someone received a score of 630. What percentage of the people taking the test score better? What percentage score worse?
- If a particular university will not admit any one scoring below 480, what percentage of the persons taking the test would be acceptable to the university?

[Delhi Univ. MBA, 2003]

Solution: (a) Given $\mu = 450$ and $\sigma = 100$. Let x be the test score. Then

$$z_1 = \frac{x - \mu}{\sigma} = \frac{500 - 450}{100} = 0.5$$

$$\text{and } z_2 = \frac{x - \mu}{\sigma} = \frac{400 - 450}{100} = -0.5$$

The area under the normal curve between $z = 0$ and $z = 0.5$ is 0.1915

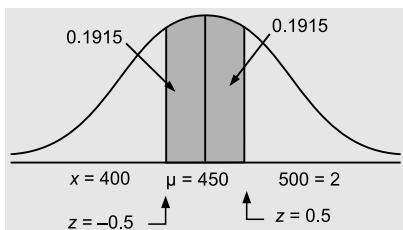
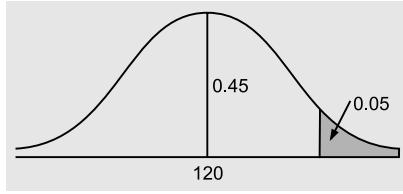
The required probability that the score falls between 400 and 500 is

$$P(400 \leq x \leq 500) = P(-0.5 \leq z \leq 0.5) = 0.1915 + 0.1915 = 0.3830$$

So the percentage of the people taking the test score between 400 and 500 is 38.30 per cent.

(b) Given $x = 630$, $\mu = 450$ and $\sigma = 100$. Thus

$$z = \frac{x - \mu}{\sigma} = \frac{630 - 450}{100} = 1.8$$



The area under the normal curve between $z = 0$ and $z = 1.8$ is 0.4641.

The probability that people taking the test score better is given by

$$P(x \geq 630) = P(z \geq 1.8) = 0.5000 + 0.4641 = 0.9640$$

That is, 96.40 percent people score better

The probability that people taking the test score worse is given by

$$P(x \leq 630) = P(z \leq 1.8) = 0.5000 - 0.4641 = 0.0359$$

That is, 3.59 per cent people score worse

(c) Given $x = 480$, $\mu = 450$ and $\sigma = 100$. Thus

$$z = \frac{x-\mu}{\sigma} = \frac{480-450}{100} = 0.30$$

The area under the normal curve between $z = 0$ and $z = 0.30$ is 0.1179. So

$$P(x \geq 480) = P(z \geq 0.30) = 0.5000 + 0.1179 = 0.6179$$

The percentage of people who score more than 480 and are acceptable to the university is 61.79 per cent.

Example 6.30: The results of particular examination are given below in a summary form:

Result	Per cent of Candidates
• Passed with distinction	10
• Passed with out distinction	60
• Failed	30

It is known that a candidate fails in the examination if he obtains less than 40 marks (out of 100) while he must obtain at least 75 marks in order to pass with distinction. Determine the mean and standard deviation of the distribution of marks, assuming this to be normal.

Solution: The given data are illustrated in the figure

Since 30 per cent candidates who obtained less than 40 marks (out of 100) failed in the examination, from the figure we have

$$z = \frac{x-\mu}{\sigma} \text{ or } -0.524 = \frac{40-\mu}{\sigma} \text{ or } \mu - 0.524\sigma = 40$$

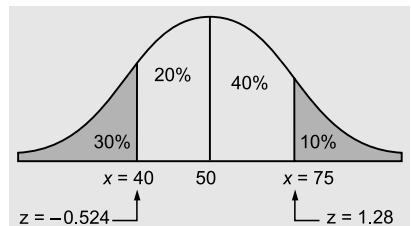
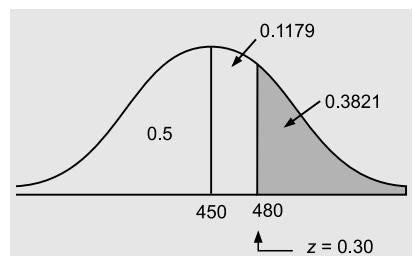
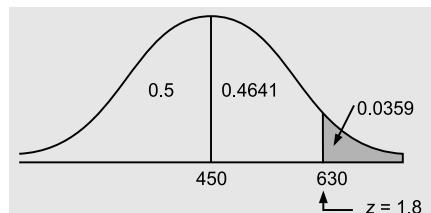
(Table value of z corresponding to 20 per cent area under the normal curve is 0.524)

Also 10 per cent candidates who obtained more than 75 marks passed with distinction, from the figure we have

$$z = \frac{x-\mu}{\sigma} \text{ or } 1.28 = \frac{75-\mu}{\sigma} \text{ or } \mu + 1.28\sigma = 75$$

(Table value of z corresponding to 40 per cent area under normal curve is 1.28)

Solving these equations, we get mean $\mu = 50.17$ and standard deviation $\sigma = 19.4$



6.6.3 Uniform (Rectangular) Distribution

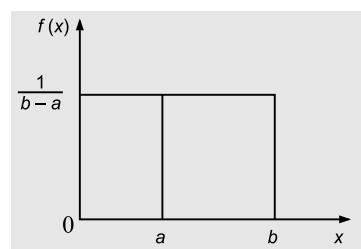
The simplest case of a continuous distribution is the uniform distribution. The general expression for the *pdf* (range of values) for a continuous random variable which is uniformly distributed over the interval between a to b is

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

This distribution is also known as *constant distribution* because the probability is constant [$= 1/(b-a)$] at every point of the interval (a, b) and is independent of whatever value the variable may take within the interval.

The general form of the rectangular probability distribution is shown in Fig. 6.9.

Figure 6.9
Rectangular Probability Distribution



The mean and variance of this distribution are given by

$$\text{Mean} = \frac{(b+a)}{2} \text{ and Variance} = \frac{(b+a)^2}{12}$$

This distribution is useful when the probability of occurrences of an event is constant whatever be the value of the variable, that is, all possible values of the continuous variable are assumed equally likely.

6.6.4 Exponential Probability Distribution

Figure 6.10
Exponential Probability Distribution

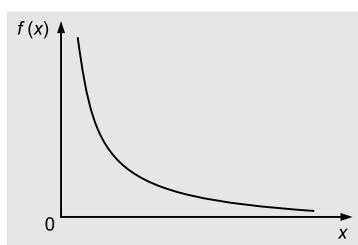
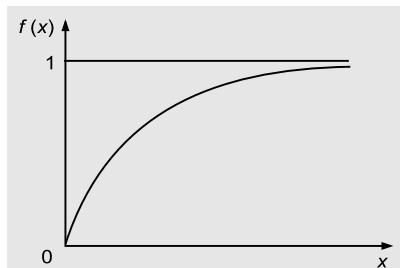


Figure 6.11
Exponential Probability Distribution



where $\mu (>0)$ is a given parameter. This distribution is also referred to as *negative exponential distribution*. It is particularly useful in the queuing (waiting line) theory.

The graph of its *pdf* slopes downward to the right from its maximum at $x = 0$, where $f(x) = \mu$, as shown in Fig. 6.10.

The exponential distribution has the mean, $1/\mu$ and variance, $1/\mu^2$. The *cumulative density function (cdf)* of the exponential distribution is

$$\begin{aligned} F(x) &= \int_0^x \mu e^{-\mu x} dx \\ &= [-e^{-\mu x}]_0^x = 1 - e^{-\mu x} \end{aligned}$$

The graph of *cdf* is shown in Fig. 6.11.

The typical applications of *cdf* of exponential functions are found in representing a saturation phenomenon. That is, the situations where the effect of successive increments of the input x (e.g., size of advertising effort) show diminishing returns (e.g., resulting sales) as the total amount of x increases, and eventually, additional input increments have no effect.

Exponential distribution is closely related with the Poisson distribution. For example, if the Poisson random variable represents the *number of arrivals* per unit time at a service window, the exponential random variable will represent the *time between two successive arrivals*.

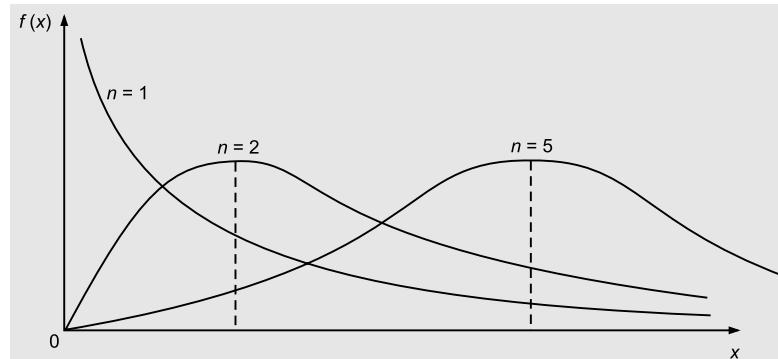
6.6.5 Gamma (or Erlang) Distribution

The probability density function (*pdf*) for the gamma (or Erlang) distribution is

$$f(x) = \frac{\mu^n (\mu x)^{n-1} e^{-\mu x}}{(n-1)!}, x > 0 \text{ and } \mu \geq 0$$

Gamma distribution is derived by the sum of n identically distributed and independent exponential random variables. Here it may be noted that the *pdf* of gamma distribution reduces to the exponential density function for $n = 1$. This means, the exponential distribution is the special case of the gamma distribution, where $n = 1$.

Figure 6.12
Gamma Distribution pdf's for $\mu = 1$



The graphs of the *pdf*'s for the gamma distribution for $\mu = 1$ and selected values of n are shown in Fig. 6.12.

In the gamma distribution *pdf*'s, the parameter μ changes the relative scales of the two axes, and the parameter n determines the location of the peak of the curve. However, for all values of these two parameters, the area under the curve is equal to 1.

The expected value and variance of this distribution are: $E(x) = n/\mu$; $\text{Var}(x) = n/\mu^2$

6.6.6 Beta Distribution

The probability density function (*pdf*) for beta distribution is

$$f(x) = \frac{x^{m-1} (1-x)^{n-1}}{\beta(m, n)} ; 0 \leq x \leq 1 ; m > 0; n > 0$$

where $\beta(m, n) = \int_0^1 x^{m-1} (1-x)^{n-1} dx$

is the beta function, whose value may be obtained directly from the table of the beta function.

The expected value and variance of random variable x in this case are given by

$$E(x) = \frac{m}{m+n} \quad \text{and} \quad \text{Var}(x) = \frac{mn}{(m+n)^2 (m+n+1)}$$

This distribution is commonly used to describe the random variable whose possible values lie in a restricted interval of numbers. A typical use of this distribution is found in the use of PERT where activity times are estimated within a specific range.

Conceptual Questions 6D

20. State the conditions under which a binomial distribution tends to (i) Poisson distribution, (ii) normal distribution. Write down the probability functions of binomial and Poisson distributions.
21. Normal distribution is symmetric with a single peak. Does this mean that all symmetric distributions are normal? Explain.
22. When finding probabilities with a normal curve we always deal with intervals; the probability of a single value of x is defined equal to zero. Why is this so?
23. When finding a normal probability, is there a difference between the values of $P(a < x < b)$ and $P(a \leq x \leq b)$, where a and b represent two numbers? Why or why not?
24. What are the parameters of normal distribution? What information is provided by these parameters?
25. What are the chief properties of normal distribution? Describe briefly the importance of normal distribution in statistical analysis. [Delhi Univ., MBA, 1990]
26. Discuss the distinctive features of the binomial, Poisson, and normal distributions. When does a binomial distribution tend to become a normal distribution? [Shukhadia Univ., MBA 1995; Kumaon Univ., MBA, 2000]
27. Briefly describe the characteristics of the normal probability distribution. Why does it occupy such a prominent place in statistics?

Self-Practice Problems 6D

- 6.40** A cigarette company wants to promote the sales of X's cigarettes (brand) with a special advertising campaign. Fifty out of every thousand cigarettes are rolled up in gold foil and randomly mixed with the regular (special king-sized, mentholated) cigarettes. The company offers to trade a new pack of cigarettes for each gold cigarette a smoker finds in a pack of brand X. What is the probability that buyers of brand X will find $X = 0, 1, 2, 3, \dots$ gold cigarettes in a single pack of 10?

- 6.41** You are in charge of rationing in a State affected by food shortage. The following reports were received from investigators:

Daily calories of food available per adult during current period

Area	Mean	S.D.
A	2000	350
B	1750	100

The estimated daily requirement of an adult is taken as

2500 calories and the absolute minimum is 1000. Comment on the reported figures and determine which area in your opinion needs more urgent attention.

- 6.42** Assume that on an average one telephone number out of fifteen is busy. What is the probability that if six randomly selected telephone numbers are called
 (a) not more than three will be busy?
 (b) at least three of them will be busy?
- 6.43** Assume the mean height of soldiers to be 68.22 inches with a variance of 10.8 inches. How many soldiers in a regiment of 1,000 would you expect to be over six feet tall?
- 6.44** The income of a group of 10,000 persons was found to be normally distributed with mean = Rs 750 p.m. and standard deviation = Rs 50. Show that in this group about 95 per cent had income exceeding Rs. 668 and only 5 per cent had income exceeding Rs 832. What was the lowest income among the richest 100?

[Delhi Univ., MBA, 1995]

- 6.45** In an intelligence test administered to 1000 students, the average score was 42 and standard deviation 24.
 Find (a) the number of students exceeding a score of 50, (b) the number of students lying between 30 and 54, (c) the value of the score exceeded by the top 100 students.
- 6.46** An aptitude test for selecting officers in a bank was conducted on 1000 candidates. The average score is 42 and the standard deviation of scores is 24. Assuming normal distribution for the scores, find:
 (a) the number of candidates whose scores exceeds 58.
 (b) the number of candidates whose scores lie between 30 and 66.

- 6.47** There are 600 business students in the postgraduate department of a university, and the probability for any student to need a copy of a particular textbook from the university library on any day is 0.05. How many

copies of the book should be kept in the university library so that the probability may be greater than 0.90 that none of the students needing a copy from the library has to come back disappointed. (Use normal approximation to the binomial probability law.)

[Kurukshetra Univ., MCom, 1998]

- 6.48** A workshop produces 2000 units of an item per day. The average weight of units is 130 kg with a standard deviation of 10 kg. Assuming normal distribution, how many units are expected to weigh less than 142 kg?

[Delhi Univ., MBA, 1996]

- 6.49** Suppose a tire manufacturer wants to set a minimum kilometer guarantee on its new AT 100 tire. Tests reveal the mean kilometer is 47,900 with a standard deviation of 2050 kms and the distribution is a normal distribution. The manufacturer wants to set the minimum guaranteed kilometer so that no more than 4 percent of the tires will have to be replaced. What minimum guaranteed kilometer should the manufacturer announce?

- 6.50** The annual commissions per salesperson employed by a pharmaceutical company, which is a manufacturer of cough syrup, averaged Rs 40,000, with a standard deviation of Rs 5000. What per cent of the salespersons earn between Rs 32,000 and Rs 42,000?

- 6.51** Management of a company is considering adopting a bonus system to increase production. One suggestion is to pay a bonus on the highest 5 per cent of production based on past experience. Past records indicate that, on the average, 4000 units of a small assembly are produced during a week. The distribution of the weekly production is approximately normal with a standard deviation of 60 units. If the bonus is paid on the upper 5 per cent of production, the bonus will be paid on how many units or more?

Hints and Answers

- 6.40** An experiment, with $n = 10$ trials, probability of finding a golden cigarette (a success) is $p = 50/100 = 0.05$. The expected number of golden cigarettes per pack is, $\lambda = np = 10(0.05)$.

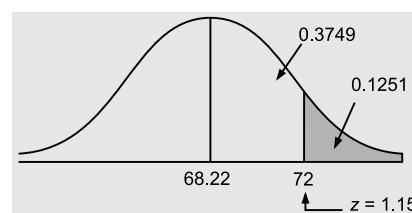
Number of Golden Cigarettes per Pack	Probability
0	0.6065
1	0.3033
2	0.0758
3	0.0126
4	0.0016

- 6.41** $\text{Area } A \quad \text{Area } B$
 Mean $\pm 3\sigma$ Mean $\pm 3\sigma$
 $= 2,000 \pm (3 \times 350)$ $= 1,750 \pm (3 \times 100)$
 or 950 and 3,050 calories 1,450 and 2,050 calories

Since the estimated requirement is minimum of 1,000 calories, area A needs more urgent attention.

- 6.43** Assuming that the distribution of height is normal. Given that, $x = 72$ inches, $\mu = 68.22$, $\sigma = \sqrt{10.8} = 3.286$. Therefore

$$z = \frac{x - \mu}{\sigma} = \frac{72 - 68.22}{3.286} = 1.15$$



Area to the right of $z = 1.15$ from the normal table is $(0.5000 - 0.3749) = 0.1251$. Probability of getting

soldiers above six feet is 0.1251 and their expected number is $0.1251 \times 1000 = 125$.

- 6.44** Given, $x = 668$, $\mu = 750$, $\sigma = 50$. Therefore

$$z = \frac{x - \mu}{\sigma} = \frac{668 - 750}{50} = -1.64$$

Area to the right of $z = -1.64$ is $(0.4495 + 0.5000) = 0.9495$.

Expected number of persons getting above Rs 668 $= 10,000 \times 0.9495 = 9495$, which is about 95 per cent of the total, that is, 10,000. Also,

$$z = \frac{832 - 750}{50} = 1.64$$

Area to the right of $z = 1.64$ is $0.5000 - 0.4495 = 0.0505$

Expected number of persons getting above Rs 832 $= 10,000 \times 0.0505 = 505$, which is approximately 5%.

Probability of getting richest 100 $= 10/1000 = 0.01$.

Value of standard normal variate for $z = 0.01$, to its right = 2.33

$$2.33 = \frac{x - 750}{50}$$

$$x = (2.33 \times 50) + 750 = \text{Rs } 866.5$$

Hence the lowest income of the richest 100 persons is Rs 866.50.

- 6.45** (a) Given $\mu = 42$, $x = 50$, $\sigma = 24$. Thus

$$z = \frac{x - \mu}{\sigma} = \frac{50 - 42}{24} = 0.333$$

Area to the right of $z = 0.333$ under the normal curve is $0.5 - 0.1304 = 0.3696$

Expected number of children exceeding a score of 50 are $0.3696 \times 1,000 = 370$.

- (b) Standard normal variate for score 30

$$z = \frac{x - \mu}{\sigma} = \frac{30 - 42}{24} = -0.5$$

Standard normal variate for score 54

$$z = \frac{x - \mu}{\sigma} = \frac{54 - 42}{24} = 0.5$$

Area between $z = 0$ to $z = 0.5 = 0.1915$

Area between $z = -0.5$ to $z = 0$ is 0.1915

Area between $z = -0.5$ to $z = 0.5$ is

$$0.1915 + 0.1915 = 0.3830$$

- 6.46** (a) Number of candidates whose score exceeds 58.

$$z = \frac{x - \mu}{\sigma} = \frac{60 - 42}{24} = 0.667$$

Area to the right of $z = 0.667$ under the normal curve is $(0.5 - 0.2476) = 0.2524$

Number of candidates whose score exceeds 60 is: $100 \times 0.2524 = 252.4$ or 252

- (b) Number of candidates whose score lies between 30 and 66.

Standard normal variate corresponding to 30,

$$z = \frac{30 - 42}{24} = -0.5$$

Standard normal variate corresponding to 66,

$$z = \frac{66 - 42}{24} = 1$$

Area between $z = -0.5$ and $z = 1$, is

$$0.1915 + 0.3413 = 0.5328$$

Number of candidates whose score lies between 30 and 66 : $1000 \times 0.5328 = 532.8$ or 533

- 6.47** Let n be the number of students and p the probability for an student to need a copy of a particular textbook from the university library. Given that $\mu = np = 600 \times 0.05 = 30$, $\sigma = \sqrt{npq} = \sqrt{600 \times 0.05 \times 0.95} = 5.34$.

Let x = number of copies of a textbook required on any day. Thus,

$$z = \frac{x - 30}{5.34} > 1.28 \text{ (95 per cent probability for } x)$$

$$x - 30 > 6.835, \text{ i.e. } x > 36.835 \approx 37 \text{ (approx.)}$$

Hence the library should keep at least 37 copies of the book to ensure that the probability is more than 90 per cent that none of the students needing a copy from the library has to come back disappointed.

- 6.48** Given $N = 2000$, $\mu = 130$, $\sigma = 10$ and $x = 142$,

$$z = \frac{x - \mu}{\sigma} = \frac{142 - 130}{10} = 1.2 \approx 0.3849$$

$$P(x \leq 142) = 0.5 + 0.3849 = 0.8849$$

Expected number of units weighing less than 142 kg is $2000 \times 0.8849 = 1,770$ approx.

- 6.49** Given $\mu = 47,900$, $\sigma = 2050$

$$P(x \leq 0.04) = P\left[z \leq \frac{x - 47,900}{2050}\right] \text{ or } -1.75 = \frac{x - 47,900}{2050}$$

$$\text{or } x = 44,312$$

The area under the normal curve to the left of μ is 0.5. So the area between x and μ is $0.5 - 0.04 = 0.46 \approx 0.4599$. This area corresponds to $z = -1.75$.

- 6.50** Given $\mu = 40$, $\sigma = 5$. Thus $P(3200 \leq x \leq 42,000) =$

$$P\left[\frac{42-40}{5} \leq z \leq \frac{32-40}{5}\right] = P[0.40 \leq z \leq -1.60] =$$

$$0.1554 + 0.4452 = 0.6006, \text{ i.e. } 60\% \text{ approx.}$$

- 6.51** $z = \frac{x - \mu}{\sigma}$ or $1.65 = \frac{x - 4000}{60}$ or $x = 4,099$ units.

Formulae Used

1. Expected value of a random variable x

$$E(x) = \sum x.P(x)$$

where x = value of the random variable

$P(x)$ = probability that the random variable will take on the value x .

2. Binomial probability distribution

- Probability of r success in n Bernoulli trials

$$P(x = r) = {}^n C_r p^r q^{n-r} = \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

where p = probability of success

q = probability of failure, $q = 1 - p$

- Mean and standard deviation of binomial distribution

$$\text{Mean } \mu = np$$

$$\text{Standard deviation } \sigma = \sqrt{npq}$$

4. Poisson probability distribution

- Probability of getting exactly r occurrences of random event

$$P(x = r) = \frac{\lambda^r e^{-\lambda}}{r!}$$

where $\lambda = np$, mean number of occurrences per interval of time

$e = 2.71828$, a constant that represents the base of the natural logarithm system

- Mean and standard deviation of Poisson distribution

$$\lambda = np, \sigma = np$$

5. Normal distribution formula:

Number of standard deviations σ a value of random variable x is away from the mean μ of normal distribution:

$$z = \frac{x - \mu}{\sigma}$$

Review Self-Practice Problems

- 6.52 Five hundred television sets are inspected as they come off the production line and the number of defects per set is recorded below:

Number of defects: 0 1 2 3 4

Number of sets: 368 72 52 7 1

Estimate the average number of defects per set and the expected frequencies of 0, 1, 2, 3, and 4 defects assuming Poisson distribution.

[Sukhadia Univ., MBA; Delhi Univ., MBA, 1997]

- 6.53 The useful life of a certain brand of radial tyre has been found to follow a normal distribution with mean $\mu = 38,000$ km and standard deviations = 3000 km. If a dealer orders 500 tyres for sale, then

- find the probability that a randomly chosen tyre will have a useful life of at least 35,000 km.
- find the approximate number of tyres that will last between 40,000 and 45,000 km.
- If an individual buys 2 tyres, then what is the probability that these tyres will last at least 38,000 km each?

- 6.54 The amount of time consumed by an individual at a bank ATM is found to be normally distributed with mean $\mu = 130$ seconds and standard deviation $\sigma = 45$ seconds.

- What is the probability that a randomly selected individual will consume less than 100 seconds at the ATM?
- What is the probability that a randomly selected individual will spend between 2 to 3 minutes at the ATM?
- Within what length of time do 20 per cent of individuals complete their job at the ATM?
- What is the least amount of time required for individuals with top 5 per cent of required time?

- 6.55 An aptitude test for selecting officers in a bank was conducted on 1000 candidates. The average score is 42 and the standard deviation of scores is 24. Assuming normal distribution for the scores, find the

- number of candidates whose scores exceed 58.

- (b) number of candidates whose scores lie between 30 and 66. [Karnataka Univ., BCom, 1995]

- 6.56 The mean inside diameter of a sample of 500 washers produced by a machine is 5.02 mm and the standard deviation is 0.05 mm. The purpose for which these washers are intended allows a maximum tolerance in the diameter of 4.96 to 5.08 mm, otherwise the washers are considered defective. Determine the percentage of defective washers produced by the machine, assuming the diameters are normally distributed.

- 6.57 In a binomial distribution consisting of 5 independent trials, the probability of 1 and 2 successes are 0.4096 and 0.2048, respectively. Find the parameter p of the distribution

- 6.58 If the probability of defective bolts be 1/10, find the following for the binomial distribution of defective bolts in a total of 400 bolts: (a) mean, (b) standard deviation, and (c) moment coefficient of skewness.

- 6.59 The probability of a bomb hitting a target is 0.20. Two bombs are enough to destroy a bridge. If 6 bombs are aimed at the bridge, find the probability that the bridge will be destroyed.

- 6.60 In an Indian university, it has been found that 25 per cent of the students come from upper income families (U), 35 per cent from middle income families (M), and 40 per cent from lower income families (L). A sample of 10 students is taken at random. What is the probability that the sample will contain 5 students from U, 2 from M and 3 from L?

- 6.61 The distribution of the total time a light bulb will burn from the time it is installed is known to be exponential with mean time between failure of the bulbs equal to 1000 hours. (a) What is the probability that a bulb will burn more than 1000 hours? and (b) what is the probability that the life will lie between 100 hours and 120 hours?

- 6.62 Past experience says that the average life of a bulb (assumed to be continuous random variable following exponential distribution) is 110 hours. Calculate the probability that the bulb will work for almost 25 hours.

[IGNOU, MS-51, 2001]

- 6.63** A firm uses a large fleet of delivery vehicles. Their record over a period of time (during which fleet size utilization may be assumed to have remained suitably constant) shows that the average number of vehicles unserviceable per day is 3. Estimate the probability on a given day when
- all vehicles will be serviceable.
 - more than 2 vehicles will be unserviceable.
- 6.64** The director, quality control of automobile company, while conducting spot checking of automatic transmission, removed ten transmissions from the pool of components and checked for manufacturing defects. In the past, only 2 per cent of the transmissions had such flaws. (Assume that flaws occur independently in different transmissions.)
- What is the probability that sample contains more than two transmissions with manufacturing flaws?
 - What is the probability that none of the selected transmission has any manufacturing flaw?
- 6.65** The Vice-President, HRD of an insurance company, has developed a new training programme that is entirely self-paced. New employees work at various stages at their own pace; completion occurs when the material is learned. The programme has been especially effective in speeding up the training process, as an employee's salary during training is only 67 per cent of that earned upon completion of the programme. In the last several years, the average completion time of the programme has been in 44 days, with a standard deviation of 12 days.
- What is the probability that an employee will finish the programme between 33 and 42 days?
- (b) What is the probability of finishing the programme in fewer than 30 days?
- 6.66** In the past 2 months, on an average, only 3 per cent of all cheques sent for clearance by a Group Housing Welfare Society (GHWS) have bounced. This month, the GHWS received 200 cheques. What is the probability that exactly ten of these cheques bounced?
- 6.67** The sales manager of an exclusive shop that sells leather clothing decides at the beginning of the winter season, how many full-length leather coats to order. These coats cost 500 each, and will sell for 1000 each. Any coat left over at the end of the season will have to be sold at a 20 per cent discount in order to make room for summer inventory. From past experience, he knows that the demand for the coats has the following probability distribution:
- | | | | | | |
|---------------------------|------|------|------|------|------|
| Number of coats demanded: | 8 | 10 | 12 | 14 | 16 |
| Probability: | 0.10 | 0.20 | 0.25 | 0.30 | 0.15 |
- He also knows that there is never any problem with selling all leftover coats at discount.
- If he decides to order 14 coats, what is the expected profit?
 - How would the answer to part (a) change if the leftover coats were sold at a 40 per cent discount?
- 6.68** Mr Tiwari is campaign manager for a candidate for Lok Sabha. General impression is that the candidate has the support of 40 per cent of registered voters. A random sample of 300 registered voters shows that 34 per cent would vote for the candidate. If 40 per cent of voters really are allied with the candidate, what is the probability that a sample of 300 voters would indicate 34 per cent or fewer on his side? Is it likely that the 40 per cent estimate is correct?

Hints and Answers

6.52	No. of defects (x) :	0	1	2	3	4
	No. of sets (f) :	368	72	52	7	1
					= 500 (= N)	
	fx :	0	72	104	21	4
					= 201 (Σfx)	

Average number of defects per set,

$$\lambda = \frac{\sum fx}{N} = \frac{201}{500} = 0.402.$$

Expected frequencies for 0, 1, 2, 3 and 4 defects are

$$\begin{aligned} NP(x=0) &= 500 e^{-\lambda} = 500 e^{-0.402} \\ &= 500 \times 0.6689 = 334.45 \end{aligned}$$

$$\begin{aligned} NP(x=1) &= NP(x=0) \times \lambda \\ &= 334.45 \times 0.402 = 134.45 \end{aligned}$$

$$\begin{aligned} NP(x=2) &= NP(x=1) \times \frac{\lambda}{2} \\ &= 134.45 \times \frac{0.402}{2} = 27.02 \end{aligned}$$

$$\begin{aligned} NP(x=3) &= NP(x=2) \times \frac{\lambda}{3} \\ &= 27.02 \times \frac{0.402}{3} = 3.62 \\ NP(x=4) &= NP(x=3) \times \frac{\lambda}{4} \\ &= 3.62 \times \frac{0.402}{4} = 3.36 \end{aligned}$$

- 6.53** (a) $z = \frac{x - \mu}{\sigma} = \frac{35,000 - 38,000}{3,000} = -1.00$
- $$\begin{aligned} P(x \geq 35,000) &= P(z \geq -1.00) \\ &= 0.500 + 0.3413 = 0.8413 \end{aligned}$$
- (b) $z_1 = \frac{x_1 - \mu}{\sigma} = \frac{40,000 - 38,000}{3,000} = 0.67$
- $$z_2 = \frac{x_2 - \mu}{\sigma} = \frac{45,000 - 38,000}{3,000} = 2.33$$
- $$\begin{aligned} P(40,000 \leq x \leq 45,000) &= P(0.67 \leq z \leq 2.33) \\ &= 0.4901 - 0.2486 = 0.2415 \end{aligned}$$

(c) $P(x \geq 38,000) = P(z \geq 0) = 0.500$

$P(2 \text{ tyres each will last at least } 38,000 \text{ km}) = (0.5000)^2 = 0.2500$ (based on the multiplication rule for the joint occurrence of independent events)

6.54 (a) $z = \frac{x - \mu}{\sigma} = \frac{100 - 130}{45} = -0.67$

$$\begin{aligned} P(x < 100) &= P(z < -0.67) = 0.5000 - 0.2486 \\ &= 0.2514 \end{aligned}$$

(b) $z_1 = \frac{120 - 130}{45} = -0.22;$

$$z_2 = \frac{180 - 130}{45} = 1.11$$

$$\begin{aligned} P(120 \leq x \leq 180) &= P(-0.22 \leq z \leq 1.11) \\ &= 0.0871 + 0.3655 = 0.4536 \end{aligned}$$

(c) $x = \mu + z\sigma = 130 + (-0.84)45 = 92 \text{ seconds}$

(d) $x = \mu + z\sigma = 130 + (1.65)45 = 204 \text{ seconds}$

6.55 (a) $z = \frac{x - \mu}{\sigma} = \frac{58 - 42}{24} = 0.67$

$$\begin{aligned} P(x > 58) &= P(z > 0.67) = 0.5000 - 0.2476 \\ &= 0.2524 \end{aligned}$$

Expected number of candidates whose score exceeds 58 is $1000 (0.2524) = 2524$

(b) $z_1 = \frac{30 - 42}{24} = -0.50; z_2 = \frac{66 - 42}{24} = 1.00$

$$\begin{aligned} P(30 \leq x \leq 66) &= P(-0.50 \leq z \leq 1.00) \\ &= 0.1915 + 0.3413 = 0.5328 \end{aligned}$$

Expected number of candidates whose scores lie between 30 and 66 is $1000 (0.5328) = 5328$.

6.56 $z_1 = \frac{x - \mu}{\sigma} = \frac{4.96 - 5.02}{0.05} = -1.20;$

$$z_2 = \frac{5.08 - 5.02}{0.05} = 1.20$$

$$\begin{aligned} P(4.96 \leq x \leq 5.08) &= P(-1.20 \leq z \leq 1.20) \\ &= 2P(0 \leq z \leq 1.20) = 2(0.3849) \\ &= 7698 \text{ or } 76.98\% \end{aligned}$$

Percentage of defective washers = $100 - 76.98 = 23.02\%$.

6.57 Given $n = 5; f(x = 1) = {}^nC_1 p^n q^{n-1}$
 $= {}^5C_1 p^1 q^4 = 5pq^4 = 0.4096$

$$f(x = 2) = {}^nC_2 p^2 q^{n-2} = {}^5C_2 p^2 q^3 = 10p^2 q^3 = 0.2048$$

$$\text{Thus } \frac{f(x = 2)}{f(x = 1)} = \frac{10p^2 q^3}{5pq^4} = \frac{0.2048}{0.4096} \text{ or } \frac{2p}{q} = \frac{1}{2}$$

or $4p = q (= 1-p)$, i.e. $p = 1/5$

6.58 Given $n = 400, p = 1/10 = 0.10, q = 9/10 = 0.90$

(a) Mean $\mu = np = 400 \times (1/10) = 40$

(b) Standard deviation

$$\sigma = \sqrt{npq} = \sqrt{400(1/10)(9/10)} = 6$$

(c) Moment coefficient of skewness

$$= \frac{q - p}{\sqrt{npq}} = \frac{0.90 - 0.10}{6} = 0.133$$

6.59 Given $p = 0.20, q = 0.80$ and $n = 6$. The bridge is destroyed if at least 2 of the bombs hit it. The required probability is

$$\begin{aligned} P(x \geq 2) &= P(x = 1) + P(x = 2) + \dots + P(x \geq 6) \\ &= 1 - [P(x = 0) + P(x = 1)] \\ &= 1 - [{}^6C_0 (0.80)^6 + {}^6C_1 (0.20) (0.80)^5] \\ &= 1 - \frac{2048}{3125} = 0.345 \end{aligned}$$

6.60 Required probability $= \frac{10!}{5! 3! 2!} (0.25)^2 (0.35)^2 (0.40)^3$
 $= 0.0193$

(Based on the rule of multinomial rule of probability)

6.61 Given, mean time between failures $1/\lambda = 1000$ or $\lambda = 1/1000$ bulbs per hours ; $t = 1000$ hours

$$\begin{aligned} (a) \quad P(t > 1000) &= 1 - P(x \leq 1000) = 1 - (1 - e^{-\lambda t}) \\ &= e^{-\lambda t} = e^{-(1/1000)1000} = e^{-1} = 0.3680 \\ (b) \quad P(100 \leq t \leq 120) &= (1 - e^{-\lambda t_1}) - (1 - e^{-\lambda t_2}) \\ &= \{1 - e^{-(1/1000)120}\} \\ &\quad \{1 - e^{-(1/1000)100}\} \\ &= 0.1132 - 0.0952 = 0.018 \end{aligned}$$

6.62 Given, mean life of a bulb $= 1/\lambda = 100$ or $\lambda = 1/100$; $t = 25$ hours

$$\begin{aligned} F(t \leq T) &= 1 - e^{-\lambda t} = 1 - e^{-(1/100)25} = 1 - e^{-0.25} \\ &= 1 - 0.7945 = 0.2055 \end{aligned}$$

$$\begin{aligned} (a) \quad P(x = 0) &= \frac{e^{-3} (3)^0}{0!} = 0.0497 \\ (b) \quad P(x > 2) &= 1 - P(x \leq 2) \\ &= 1 - [P(x = 0) + P(x = 1) + P(x = 2)] \\ &= 1 - \left[e^{-3} + 3e^{-3} + \frac{9}{2} e^{-3} \right] \\ &= 1 - e^{-3} \left(1 + 3 + \frac{9}{2} \right) = 1 - \frac{11}{2} (0.0497) \\ &= 1 - 0.4224 = 0.5776 \end{aligned}$$

6.64 (a) Given, $p = 0.02, q = 0.98, n = 10$

$$\begin{aligned} P(x > 2 \text{ flaws}) &= 1 - [P(x = 0) + P(x = 1) + P(x = 2)] \\ &= 1 - [{}^{10}C_0 (0.98)^{10} + {}^{10}C_1 (0.02) \\ &\quad (0.98)^9 + {}^{10}C_2 (0.02)^2 (0.98)^8] \\ &= 1 - [0.8171 + 0.1667 + 0.0153] \\ &= 0.0009 \end{aligned}$$

$$(b) \quad P(x = 0 \text{ flaw}) = {}^nC_0 p^0 q^n = {}^{10}C_0 (0.02)^0 (0.98)^{10}$$

$$= 10 (0.98)^{10} = 0.8171$$

6.65 Given, average completion time of the programme, $\mu = 44$ days and standard deviation, $\sigma = 12$ days

$$\begin{aligned} (a) \quad P(33 \leq x \leq 42) &= P\left[\frac{x_1 - \mu}{\sigma} \leq z \leq \frac{x_2 - \mu}{\sigma}\right] \\ &= P\left[\frac{33 - 44}{12} \leq z \leq \frac{42 - 44}{12}\right] \\ &= P[-0.92 \leq z \leq -0.17] \\ &= 0.3212 - 0.0675 = 0.2537 \end{aligned}$$

$$\begin{aligned} (b) \quad P(x < 30) &= P\left[z < \frac{x - \mu}{\sigma}\right] = P\left[z < \frac{30 - 44}{12}\right] \\ &= P(z < -1.7) = 0.5000 - 0.3790 \\ &= 0.1210 \end{aligned}$$

6.66 Given $n = 200$, $p = 0.03$, $\lambda = np = 200(0.03) = 6$.

$$P(x=10) = \frac{e^{-\lambda} \lambda^r}{r!} = \frac{e^{-6} (6)^{10}}{10!} = 0.0413$$

6.67 Expected profit when 14 coats are ordered

Number of coats				
ordered	:	8	10	12
Probability	:	0.10	0.20	0.25
(a) Profit per coat (Rs)	:	1160	1240	1320
Total expected profit (Rs)	:	116	248	330
(b) Profit per Coat (Rs)	:	920	1080	1240

Total expected

$$\text{profit (Rs): } 92 \quad 216 \quad 310 \quad 620 = 1248$$

6.68 Given $n=300$, $p=0.40$; $\mu=np=300(0.40)=120$;

$$\sigma = \sqrt{npq} = \sqrt{120(0.06)} = 8.48$$

$$P(x \leq 0.34 \times 300 = 102)$$

$$= P\left[z < \frac{x - \mu}{\sigma}\right] = P\left[z < \frac{102 - 120}{8.48}\right]$$

$$= P[z \leq -2.12] = 0.5000 - 0.4830$$

$$= 0.0170$$

Since the probability that the sample would indicate 34 per cent or less is very small, it is unlikely that the 40 per cent estimate is correct.

This page is intentionally left blank.

By a small sample we may judge of the whole piece.

—Cervantes

Nine times out of ten, in the arts as in life, there is actually no truth to be discovered; there is only error to be exposed.

— H. L. Mencken

Sampling and Sampling Distributions

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- distinguish between population parameter and sample statistics
- apply the Central Limit Theorem
- know various procedures of sampling that provide an attractive means of learning about a population or process.
- develop the concept of a sampling distribution that helps you understand the methods and underlaying thinking of statistical inference.

7.1 INTRODUCTION

So far we introduced certain statistical methods (Chapters 2 to 4) to analyse a data set and the concepts of probability and its distributions (Chapters 5 and 6) to increase our knowledge about unknown features (characteristics) of a population or a process. In statistical inference we use random sample or samples to extract information about the population from which it is drawn. The information we extract is in the form of summary statistics: a sample mean, a sample standard deviation or other measures computed from the sample. Sample statistics are treated as estimator of population parameters – μ , σ , p , etc.

Sampling The process of selecting a sample from a population is called *sampling*. In sampling, a representative *sample* or *portion* of elements of a population or process is selected and then analysed. Based on sample results, called *sample statistics*, *statistical inferences* are made about the population characteristic. For instance, a political analyst selects specific or random set of people for interviews to estimate the proportion of the votes that each candidate may get from the population of voters; an auditor selects a sample of vouchers and calculates the sample mean for estimating population average amount; or a doctor examines a few drops of blood to draw conclusions about the nature of disease or blood constitution of the whole body.

7.2 REASONS OF SAMPLE SURVEY

A census is a count of all the elements in a population. Few examples of census are: population of eligible voters; census of consumer preference to a particular product, buying habits of adult Indians. Some of the reasons to prefer sample survey instead of census are given below.

1. **Movement of Population Element** The population of fish, birds, snakes, mosquitoes, etc. are large and are constantly moving, being born and dying. So instead of attempting to count all elements of such populations, it is desirable to make estimates using techniques such as counting birds at a place picked at random, setting nets at predetermined places, etc.
2. **Cost and/or Time Required to Contact the Whole Population** A census involves a complete count of every individual member of the population of interest, such as persons in a state, households in a town, shops in a city, students in a college, and so on. Apart from the cost and the large amount of resources (such as enumerators, clerical assistance, etc.) that are required, the main problem is the time required to process the data. Hence the results are known after a big gap of time.
3. **Destructive Nature of Certain Tests** The census becomes extremely difficult, if not impossible, when the population of interest is either infinite in terms of size (number); constantly changing; in a state of movement; or observation results require destruction. For example, sometimes it is required to test the strength of some manufactured item by applying a stress until the unit breaks. The amount of stress that results in breakage is the value of the observation that is recorded. If this procedure is applied to an entire population, there would be nothing left. This type of testing is called destructive testing and requires that a sample be used in such cases.

7.3 TYPES OF BIAS DURING SAMPLE SURVEY

Not all surveys produce trust worthy results. Results based on a survey are *biased* when the method used to obtain those results would consistently produce values that are either too high or too low. Following are the common types of bias that might occur in surveys:

1. **Undercoverage Bias** This bias occurs when a random sample chosen does not represent the population of interest. For instance, passengers at a railway station are surveyed to determine attitude towards buying station ticket, the results are not likely to represent all passengers at railway stations.
2. **Non-response Bias** This bias occurs when only a small number of respondents respond or return their questionnaire. The sample results would be biased because only those respondent who were particularly concerned about the subject chose to respond.
3. **Wording Bias** This bias occurs when respondents respond differently from how they truly feel. It may be due to the reason that the questionnaire contains questions that tend to confuse the respondents. For instance, questions for the survey about the use of drug, payment of income tax, abusive behavior, etc. must be worded and conducted carefully to minimise response bias.

Sampling error The absolute value of the difference between an unbiased estimate and the corresponding population parameter, such as $|\bar{x} - \mu|$, $|\bar{p} - p|$, etc.

7.3.1 Sampling and Non-Sampling Errors

Any statistical inference based on sample results (statistics) may not always be correct, because sample results are either based on partial or incomplete analysis of the population features (or characteristics). This error is referred to as the **sampling error** because each sample taken may produce a different estimate of the population characteristic compared to those results that would have been obtained by a complete enumeration of the population. It is, therefore, necessary to measure these errors so as to have an exact idea about the reliability of sample-based estimates.

of population features. The likelihood that a sampling error exceeds any specified magnitude must always be specified in terms of a probability value, say 5%. This acceptable margin of error is then used to produce a *confidence* in the decision-maker to arrive at certain conclusions with the limited data at his disposal. In general, in the business context, decision-makers wish to be 95 per cent or more confident that the range of values of sample results reflect the true characteristic of the population or process of interest.

Non-sampling errors arise during census as well as sampling surveys due to biases and mistakes such as (i) incorrect enumeration of population members, (ii) non-random selection of samples, (iii) use of incomplete, vague, or faulty questionnaire for data collection, or (iv) wrong editing, coding, and presenting of the responses received through the questionnaire. The sampling errors can be minimized if (i) the questionnaire contains precise and unambiguous questions, (ii) the questionnaire is administered carefully, (iii) the interviewers are given proper training, and (iv) the responses are correctly processed.

Measurement of Sampling Error A measure of sampling error is provided by the standard error of the estimate. Estimation of sampling error can reduce the element of uncertainty associated with interpretation of data. In most cases, the degree of precision or the level of error, would depend on the size of the sample.

The standard error of estimate is inversely proportional to the square root of the sample size. In other words, as the sample size increases, element of error is reduced. Figure 7.1 illustrates this concept.

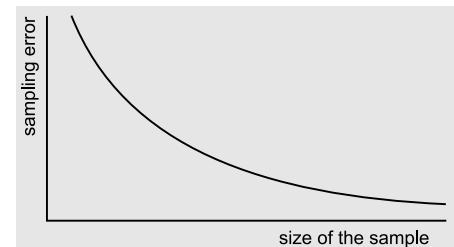


Figure 7.1
Measurement of Sampling Error

7.4 POPULATION PARAMETERS AND SAMPLE STATISTICS

Parameters An exact, but generally unknown measure (or value) which describes the entire population or process characteristics is called a *parameter*. For example, quantities such as mean μ , variance σ^2 , standard deviation σ , median, mode, and proportion p computed from a data set (also called population) are called parameters. A parameter is usually denoted with letters of the lower case Greek alphabet, such as mean μ and standard deviation σ .

Sample Statistics A measure (or value) found from analysing sample data is called a *sample statistic* or simply a *statistic*. Inferential statistical methods attempt to estimate population parameters using sample statistics. **Sample statistics** are usually denoted by Roman letters such as mean \bar{x} , standard deviation s , variance s^2 and proportion \bar{p} .

The value of every statistic varies randomly from one sample to another whereas the value of a parameter is considered as constant. The value for statistic calculated from any sample depends on the particular random sample drawn from a population. Thus probabilities are attached to possible outcomes in order to assess the reliability or sample error associated with a statistical inference about a population based on a sample. Figure 7.2 shows the estimation relationships between sample statistics and the population parameters.

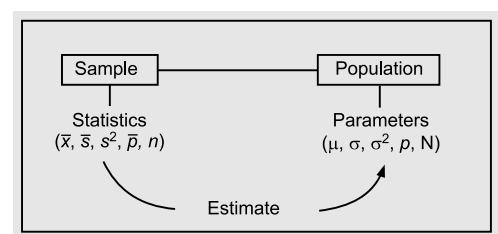


Figure 7.2
Estimation Relationship between Sample and Population Measures

7.5 PRINCIPLES OF SAMPLING

The following are two important principles which determine the possibility of arriving at a valid statistical inference about the features of a population or process:

- (i) Principle of statistical regularity
- (ii) Principle of inertia of large numbers

Sample statistic A sample measure, such as mean \bar{x} , standard deviation, s , proportion \bar{p} , and so on.

7.5.1 Principle of Statistical Regularity

This principle is based on the mathematical theory of probability. According to King, ‘*The law of statistical regularity lays down that a moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristic of the large group.*’ This principle, emphasises on two factors:

- (i) **Sample Size Should be Large** As the size of sample increases, it becomes more and more representative of parent population and shows its characteristics. However, in actual practice, large samples are more expansive. Thus a balance has to be maintained between the sample size, degree of accuracy desired and financial resources available.
- (ii) **Samples Must be Drawn Randomly** The random sample is the one in which elements of the population are drawn in a such way that each combination of elements has an equal probability of being selected in the sample. When the term random sample is used without any specification, it usually refers to a *simple random sample*. The selection of samples based on this principle can reduce the amount of efforts required in arriving at a conclusion about the characteristic of a large population. For example, to understand the book buying habit of students in a college, instead of approaching every student, it is easy to talk to a randomly selected group of students to draw the inference about all students in the college.

7.5.2 Principle of Inertia of Large Numbers

This principle is a corollary of the principle of statistical regularity and plays a significant role in the sampling theory. This principle states that, under similar conditions, *as the sample size (number of observations in a sample) get large enough, the statistical inference is likely to be more accurate and stable.* For example, if a coin is tossed a large number of times, then relative frequency of occurrence of head and tail is expected to be equal.

7.6 SAMPLING METHODS

As mentioned above, sampling methods compared to census provides an attractive means of learning about a population or process in terms of reduced cost, time and greater accuracy. The representation basis and the element selection techniques from the given population, classify several sampling methods into two categories as shown in Tabel 7.1.

Table 7.1: Types of Sampling Methods

Element Selection	Representation Basis	
	Probability (Random)	Non-probability (Non-random)
• Unrestricted	Simple random sampling	Convenience sampling
• Restricted	Complex random sampling • Stratified sampling • Cluster sampling • Systematic sampling • Multi-stage sampling	Purposive sampling • Quota sampling • Judgement sampling

7.6.1 Probability Sampling Methods

Several probability sampling methods for selecting samples from a population or process are as follows:

Simple Random (Unrestricted) Sampling In this method, every member (or element) of the population has an equal and independent chance of being selected again and

again when a sample is drawn from the population. To draw a random sample, we need a complete list of all elements in the population of interest so that each element can be identified by a distinct number. Such a list is called *frame for experiment*. The frame for experiment allows us to draw elements from the population by randomly generating the numbers of the elements to be included in the sample.

For instance, in drawing the random sample of 50 students from a population of 3500 students in a college we make a list of all 3500 students and assign each student an identification number. This gives us a list of 3500 numbers, called frame for experiment. Then we generate by computer or by other means a set of 50 random numbers in the range of values from 1 and 3500. The procedure gives every set of 50 students in the population an equal chance of being included in the sample. Selecting a random sample is analogous to using a gambling device to generate numbers from this list.

This method is suitable for sampling, as many statistical tests assume independence of sample elements. One disadvantage with this method is that all elements of the population have to be available for selection, which many a times is not possible.

Stratified Sampling This method is useful when the population consists of a number of heterogeneous subpopulations and the elements within a given subpopulation are relatively homogeneous compared to the population as a whole. Thus, population is divided into mutually exclusive groups called *strata* that are relevant, appropriate and meaningful in the context of the study. A simple random sample, called a *sub-sample*, is then drawn from each *strata* or *group*, in proportion or a non-proportion to its size. As the name implies, a proportional sampling procedure requires that the number of elements in each stratum be in the same proportion as in the population. In non-proportional procedure, the number of elements in each stratum are disproportionate to the respective numbers in the population. The basis for forming the strata such as location, age, industry type, gross sales, or number of employees, is at the discretion of the investigator. Individual stratum samples are combined into one to obtain an overall sample for analysis.

This sampling procedure is more efficient than the simple random sampling procedure because, for the same sample size, we get more representativeness from each important segment of the population and obtain more valuable and differentiated information with respect to each strata. For instance, if the president of a company is concerned about low motivational levels or high absentee rate among the employees, it makes sense to stratify the population of organizational members according to their job levels. Assume that the 750 employees were divided into six strata as shown in Table 7.2. Let 100 employees are to be selected for study, then number to be selected from each strata is shown in Table 7.2

Table 7.2: Proportionate and Disproportionate Stratified Random Samples

Strata	Job Level	Number of Employees (Elements)	Number of Employees in the Sample	
			Proportionate Sample	Disproportionate Sample
1	Top management	15	$(15/750) \times 100 = 2$	3
2	Middle-level management	30	$(30/750) \times 100 = 4$	10
3	Lower-level management	55	$(55/750) \times 100 = 7$	15
4	Supervisors	105	$(105/750) \times 100 = 14$	25
5	Clerks	510	$(510/750) \times 100 = 68$	37
6	Secretaries	35	$(35/750) \times 100 = 5$	10
		750	100	100

When the data are collected and the analysis completed, it is likely that the members of a particular group are found to be not motivated. This information will help in taking action at the right level and think of better ways to motivate that group members which otherwise would not have been possible.

Disproportionate sampling decisions are made either when strata are either too small, too large, or when there is more variability suspected within a particular stratum. For example, the educational levels in a particular strata might be expected to influence perceptions, so more people will be sampled at this level. Disproportionate sampling is done when it is easier, and less expensive to collect data from one or more strata than from others.

For this method of sampling to be more effective in terms of reliability, efficiency, and precision, any stratification should be done which ensures

- (i) maximum uniformity among members of each strata,
- (ii) largest degree of variability among various strata.

Cluster Sampling This method, sometimes known as *area sampling method*, has been devised to meet the problem of costs or inadequate sampling frames (a complete listing of all elements in the population so that each member can be identified by a distinct number). The entire population to be analysed is divided into smaller groups or chunks of elements and a sample of the desired number of areas selected by a simple random sampling method. Such groups are termed as *clusters*. The elements of a cluster are called *elementary units*. These clusters do not have much heterogeneity among the elements. A household where individuals live together is an example of a cluster.

If several groups with intragroup heterogeneity and intergroup homogeneity are found, then a random sampling of the clusters or groups can be done with information gathered from each of the elements in the randomly chosen clusters. Cluster samples offer more heterogeneity within groups and more homogeneity among groups—the reverse of what we find in stratified random sampling, where there is homogeneity within each group and heterogeneity across groups.

For instance, committees formed from various departments in an organization to offer inputs to make decisions on product development, budget allocations, marketing strategies, etc are examples of different clusters. Each of these clusters or groups contains a heterogeneous collection of members with different interests, orientations, values, philosophy, and vested interests. Based on individual and combined perceptions, it is possible to make final decision on strategic moves for the organization.

In summary, cluster sampling involves preparing only a list of clusters instead of a list of individual elements. For examples, (i) residential blocks (colonies) are commonly used to cluster in surveys that require door-to-door interviews, (ii) airlines sometimes select randomly a set of flights to distribute questionnaire to every passenger on those flights to measure customer satisfaction. In this situation, each flight is a cluster. It is much easier for the airline to choose a random sample of flights than to identify and locate a random sample of individual passengers to distribute questionnaire.

Multistage Sampling This method of sampling is useful when the population is very widely spread and random sampling is not possible. The researcher might stratify the population in different regions of the country, then stratify by urban and rural and then choose a random sample of communities within these strata. These communities are then divided into city areas as clusters and randomly consider some of these for study. Each element in the selected cluster may be contacted for desired information.

For example, for the purpose of a national pre-election opinion poll, the *first stage* would be to choose as a sample a specific state (region). The size of the sample, that is the number of interviews, from each region would be determined by the relative populations in each region. In the *second stage*, a limited number of towns/cities in each of the regions would be selected, and then in the *third stage*, within the selected towns/cities, a sample of respondents could be drawn from the electoral roll of the town/city selected at the second stage.

The essence of this type of sampling is that a subsample is taken from successive groups or strata. The selection of the sampling units at each stage may be achieved with or without stratification. For example, at the second stage when the sample of towns/cities is being drawn, it is customary to classify all the urban areas in the region in such

a way that the elements (towns/cities) of the population in those areas are given equal chances of inclusion.

Systematic Sampling This procedure is useful when elements of the population are already physically arranged in some order, such as an alphabetized list of people with driving licenses, list of bank customers by account numbers. In these cases one element is chosen at random from first k element and then every k th element (member) is included in the sample. The value k is called the *sampling interval*. For example, suppose a sample size of 50 is desired from a population consisting of 100 accounts receivable. The sampling interval is $k = N/n = 1000/50 = 20$. Thus a sample of 50 accounts is identified by moving systematically through the population and identifying every 20th account after the first randomly selected account number.

7.6.2 Non-Random Sampling Methods

Several non-random sampling methods for selecting samples from a population or process are as follows:

Convenience Sampling In this procedure, units to be included in the sample are selected at the convenience of the investigator rather than by any prespecified or known probabilities of being selected. For example, a student for his project on 'food habits among adults' may use his own friends in the college to constitute a sample simply because they are readily available and will participate for little or no cost. Other examples are, public opinion surveys conducted by any TV channel near the railway station; bus stop, or in a market.

Convenience samples are easy for collecting data on a particular issue. However, it is not possible to evaluate its representativeness of the population and hence precautions should be taken in interpreting the results of convenient samples that are used to make inferences about a population.

Purposive Sampling Instead of obtaining information from those who are most conveniently available, it sometimes becomes necessary to obtain information from specific targets–respondents who will be able to provide the desired information either because they are the only ones who can give the desired information or because they satisfy to some criteria set by researcher.

Judgement Sampling Judgement sampling involves the selection of respondents who are in the best position to provide the desired information. The judgment sampling is used when a limited number of respondents have the information that is needed. In such cases, any type of probability sampling across a cross section of respondents is purposeless and not useful. This sampling method may curtail the generalizability of the findings due to the fact that we are using a sample of respondents who are conveniently available to us. It is the only viable sampling method for obtaining the type of information that is required from very specific section of respondents who possess the knowledge and can give the desired information.

However, the validity of the sample results depend on the proper judgment of the investigator in choosing the sample. Great precaution is needed in drawing conclusions based on judgment samples to make inferences about a population.

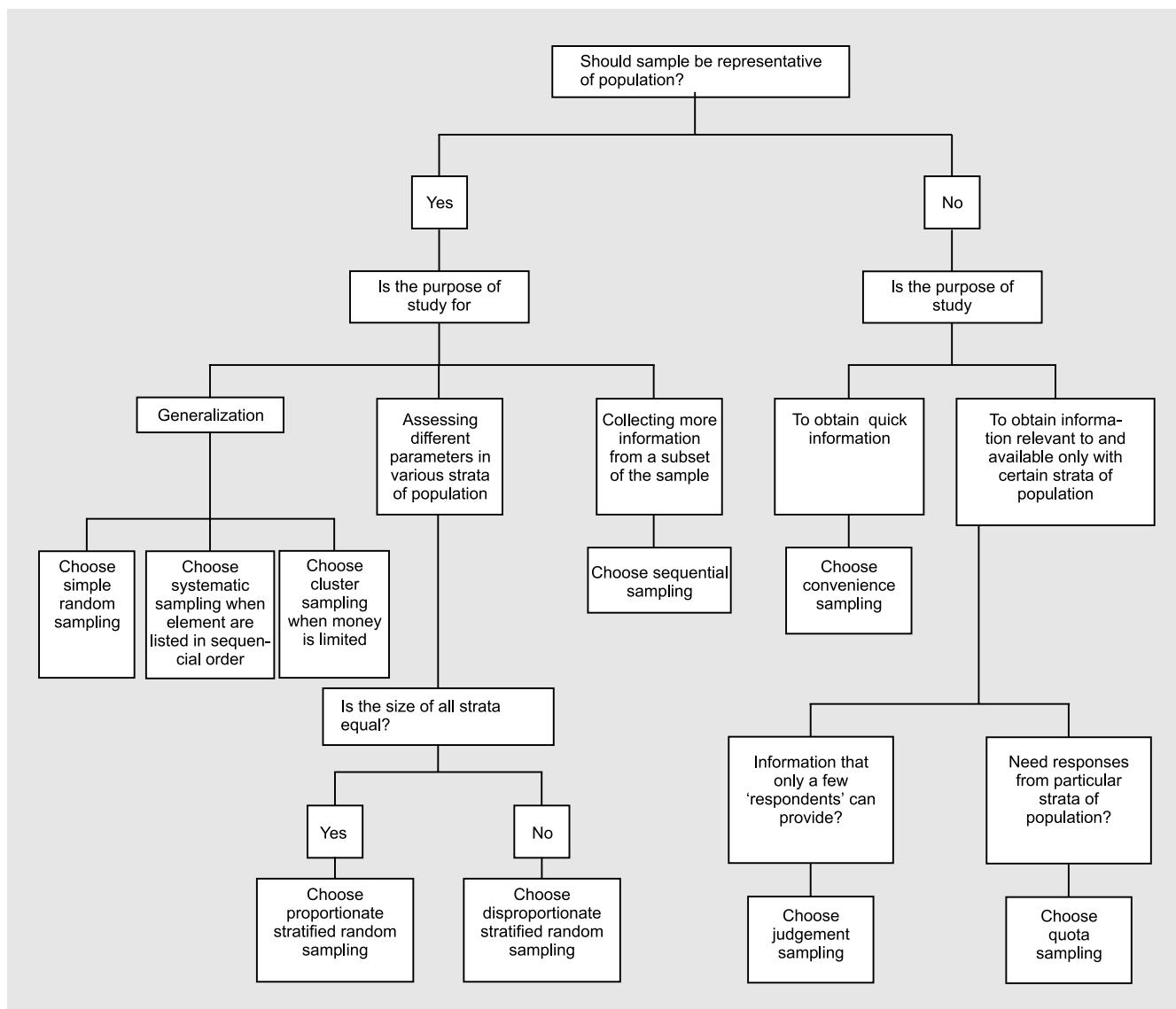
Quota Sampling Quota Sampling is a form of proportionate stratified sampling in which a predetermined proportion of elements are sampled from different groups in the population, but on convenience basis. In other words, in quota sampling the selection of respondents lies with the investigator, although in making such selection he/she must ensure that each respondent satisfies certain criteria which is essential for the study. For example, the investigator may choose to interview ten men and ten women in such a way that two of them have annual income of more than two lakh rupees five of them have annual income between one and two lakh rupees and thirteen whose annual income is below one lakh rupees. Furthermore, some of them should be between 25 and 35 years of age, others between 36 and 45 years of age, and the balance over 45 years. This means that the investigator's choice of respondent is partly dictated by these 'controls'.

Quota sampling has been criticized because it does not satisfy the fundamental requirement of a sample, that is, it should be random. Consequently, it is not possible to achieve precision of results on any valid basis.

7.6.3 Choice of Sampling Methods

Figure 7.3
Guidelines to Choose Sample

The choice of particular sampling method (procedure) must be decided according to various factors such as: nature of study, size of the population, size of the sample, availability of resources, degree of precision desired, etc. A choice plan is shown in Fig. 7.3.



Judging the Reliability of a Sample The reliability of a sample can be determined in the following ways to ensure dependable results:

- A number of samples may be taken from the same population and the results of various samples compared. If there is not much variation in the results of the different samples, it is a measure of its reliability.
- Sub-sample may be taken from the main sample and studied. If the results of the sub-samples are similar to those given by the main sample it gives a measure of its reliability.
- If some mathematical properties are found in the distribution under study, the sample result can be compared with expected values obtained on the

basis of mathematical relationship and if the difference between them is not significant, the sample has given dependable results.

In probability distributions where binomial, normal, Poisson or any other theoretical probability distribution is applicable, sample results can be compared with the expected values to get an idea about the reliability of the sample.

7.7 SAMPLING DISTRIBUTIONS

In Chapter 3 we have discussed several statistical methods to calculate parameters such as the mean and standard deviation of the population of interest. These values were used to describe the characteristics of the population. If a population is very large and the description of its characteristics is not possible by the census method, then to arrive at the statistical inference, samples of a given size are drawn repeatedly from the population and a particular '*statistic*' is computed for each sample. The computed value of a particular statistic will differ from sample to sample. In other words, if the same statistic is computed for each of the samples, the value is likely to vary from sample to sample. Thus, theoretically it would be possible to construct a frequency table showing the values assumed by the statistic and the frequency of their occurrence. This *distribution of values of a statistic* is called a **sampling distribution**, because the values are the outcome of a process of sampling. Since the values of statistic are the result of several simple random samples, therefore these are random variables.

Suppose all possible random samples of size n are drawn from a population of size N , and the 'mean' values computed. This process will generate a set of ${}^N C_n = N!/n!(N-n)!$ sample means, which can be arranged in the form of a distribution. This distribution would have its mean denoted by $\mu_{\bar{x}}$ and standard deviation is denoted by $\sigma_{\bar{x}}$ (also called *standard error*). We may follow this procedure to compute any other statistic from all possible samples of given size drawn from a population.

The concept of sampling distribution can be related to the various probability distributions. Probability distributions are the theoretical distributions of random variables that are used to describe characteristics of populations or processes under certain specified conditions. That is, probability distributions are helpful in determining the probabilities of outcomes of random variables when populations or processes that generate these outcomes satisfy certain conditions. For example, if a population has a normal distribution, then the phenomenon that describes the normal probability distribution provides a useful description of the distribution of population values. Thus when mean values obtained from samples are distributed normally, it implies that this distribution is useful for describing the characteristics (or properties) of sampling distribution. Consequently, these properties, which are also the properties of sampling distribution, help to frame rules for making statistical inferences about a population on the basis of a single sample drawn from it, that is, without even repeating the sampling process. The sampling distribution of a sample statistic also helps in describing the extent of error associated with an estimate of the value of population parameters.

7.7.1 Standard Error of Statistic

Since sampling distribution describes how values of a sample statistic, say mean, is scattered around its own mean $\mu_{\bar{x}}$, therefore its standard deviation $\sigma_{\bar{x}}$ is called the *standard error* to distinguish it from the standard deviation σ of a population. The population standard deviation describes the variation among values of the members of the population, whereas the standard deviation of sampling distribution measures the variability among values of the sample statistic (such as mean values, proportion values) due to sampling errors. Thus knowledge of sampling distribution of a sample statistic enables us to determine the probability of sampling error of the given magnitude. Consequently standard deviation of sampling distribution of a sample statistic measures sampling error and is also known as *standard error of statistic*.

Sampling distribution A probability distribution consisting of all possible values of a sample statistic.

The standard error of statistic measures not only the amount of chance error in the sampling process but also the accuracy desired. One of the most common inferential procedures is *estimation*. In estimation, the value of the statistic is used as an estimate of the value of the population parameter.

7.7.2 Distinction between Population, Sample Distributions, and Sampling Distributions

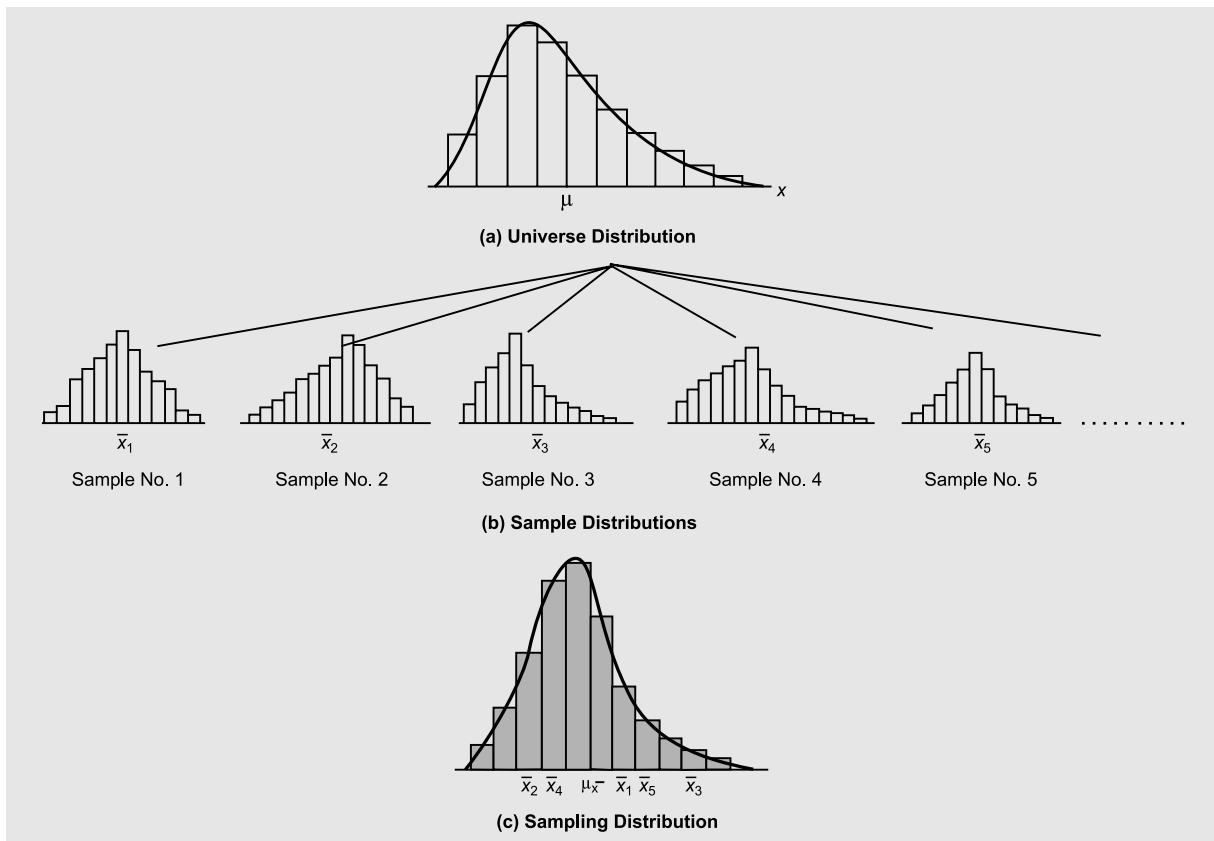
In the previous sections, we introduced sampling methods to drawn samples of the same size repeatedly from a population, and compute for each sample the statistic of interest. Hence from the distribution of population we can derive a *sampling distribution of statistic of interest*. This distribution has its own mean and standard deviation (also called *standard error*). Such distributions describe the relative frequency of occurrence of values of a sample statistic and hence help to estimate universal parameters.

Population distribution is the distribution of values of its elements members and has mean denoted by μ , variance σ^2 and standard deviation σ .

Sample distribution is the distribution of measured values of statistic in random samples drawn from a given population. Each sample distribution is a discrete distribution [as shown in Fig 7.4(b)] because the value of the sample mean would vary from sample to sample. This variability serves as the basis for the random sampling distribution. In Fig. 7.4(b) only five such samples are shown, however, there could be several such cases. In such distributions the arithmetic mean represents the average of all possible sample means or the ‘mean of means’ denoted by \bar{x} ; the standard deviation which measures the variability among all possible values of the sample values, is considered as a good approximation of the population’s standard deviations σ . To estimate σ of the population to greater accuracy the formula

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}} \text{ is used instead of } s = \sqrt{\frac{\sum(x - \bar{x})^2}{n}}$$

Figure 7.4
Random Sampling Distribution of Sample Mean



where n is the size of sample. The new value $n - 1$ in the denominator results into higher value of s than the observed value s of the sample. Here $n - 1$ is also known as *degree of freedom*. The number of degrees of freedom, $df = n - 1$ indicate the number of values that are free to vary in a random sample.

Sampling distribution is the distribution of all possible values of a statistic from all the distinct possible samples of equal size drawn from a population or a process as shown in Fig. 7.4(c). The sampling distribution of the mean values has its own arithmetic mean denoted by $\mu_{\bar{x}}$ (mu sub x bar) or \bar{x} (mean of mean values) and standard deviation $\sigma_{\bar{x}}$ (sigma sub x bar) or s . The standard deviation of the sampling distribution indicates how different samples would be distributed. The calculation of these sampling distribution statistics is based on the following properties:

- The arithmetic mean $\mu_{\bar{x}}$ of sampling distribution of mean values is equal to the population mean μ regardless of the form of population distribution, that is, $\mu_{\bar{x}} = \mu$.
- The sampling distribution has a standard deviation (also called standard error or sampling error) equal to the population standard deviation divided by the square root of the sample size, that is, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.
- Remember that a standard deviation is the spread of the values around the average in a *single sample*, whereas the standard error is the spread of the averages around the average of averages in a *sampling distribution*.
- The sampling distribution of sample mean values from normally distributed populations is the normal distribution for samples of all sizes.

Sampling error provides some idea of the precision of a statistical estimate. A low sampling error means a relatively less variability or range in the sampling distribution. Since we never actually see the sampling distribution, calculation of sampling error is based on the calculation of the standard deviation of the sample. Thus greater the sample's standard deviation, the greater the standard error (and the sampling error). The standard error is also related to the sample size: greater the sample size, smaller the standard error.

A sample of size $n \geq 30$ is generally considered to be a large sample for statistical analysis whereas a sample of size $n < 30$ is considered to be a small sample. It may be noted from the formula of $\sigma_{\bar{x}}$ that its value tends to be smaller as the size of sample n increases and vice-versa.

When standard deviation σ of population is not known, the standard deviation s of the sample, which closely approximates σ value, is used to compute standard error, that is, $\sigma_{\bar{x}} = s/\sqrt{n}$.

Conceptual Questions 7A

- Briefly explain
 - The fundamental reason for sampling
 - Some of the reasons why a sample is chosen instead of testing the entire population
- What is the relationship between the population mean, the mean of a sample, and the mean of the distribution of the sample mean?
- Is it possible to develop a sampling distribution for other statistics besides sample mean? Explain.
- How does the standard error of mean measure sampling error? Is the amount of sampling error in the sample mean affected by the amount of variability in the universe? Explain.
- If only one sample is selected in a sampling problem,
- how is it possible to have an entire distribution of the sample mean?
- What is sampling? Explain the importance in solving business problems. Critically examine the well-known methods of probability sampling and non-probability sampling. [Delhi Univ., MBA, 1998]
- Point out the differences between a sample survey and a census survey. Under what conditions are these undertaken? Explain the law which forms the basis of sampling. [Delhi Univ., MBA, 1999]
- Explain with the help of an example, the concept of sampling distribution of a sample statistic and point out its role in managerial decision-making. [Delhi Univ., MBA, 2000]

9. Why does the sampling distribution of mean follow a normal distribution for a large sample size even though the population may not be normally distributed?
10. Explain the concept of standard error. Discuss the role of standard error in large sample theory.
11. What do you mean by sampling distribution of a statistic and its standard error? Give the expressions for the standard error of the sample mean.
12. Bring out the importance of sampling distribution and the concept of standard error in statistical application.
13. Explain the principles of 'Inertia of Large Numbers' and 'Statistical Regularity'.
14. Enumerate the various methods of sampling and describe two of them mentioning the situations where each one is to be used.
15. Distinguish between sampling and non-sampling errors.
- What are their sources? How can these errors be controlled?
16. (a) What is the distinction between a sampling distribution and a probability distribution?
(b) What is the distinction between a standard deviation and a standard error?
17. Is the standard deviation of sampling distribution of mean the same as the standard deviation of the population? Explain.
18. Explain the terms 'population' and 'sample'. Explain, why is it sometimes necessary and often desirable to collect information about the population by conducting a sample survey instead of complete enumeration?
19. What are the main steps involved in a sample survey. Discuss different sources of error in such surveys and point out how these errors can be controlled.

7.8 SAMPLING DISTRIBUTION OF SAMPLE MEAN

In general, the sampling distribution of sample means depending on the distribution of the population or process from which samples are drawn. If a population or process is normally distributed, then sampling distribution of sample means is also normally distributed regardless of the sample size. Even if the population or process is not distributed normally, the sampling distribution of sample mean tends to be distributed normally as the sample size is sufficiently large.

7.8.1 Sampling Distribution of Mean When Population has Non-Normal Distribution

If population is not normally distributed, then we make use of the **central limit theorem** to describe the random nature of the sample mean for large samples without knowledge of the population distribution. The Central Limit Theorem states that

- When the random samples of observations are drawn from a non-normal population with finite mean μ and standard deviation σ , and as the sample size n is increased, the sampling distribution of sample mean \bar{x} is approximately normally distributed, with mean and standard deviation as:

$$\mu_{\bar{x}} = \mu \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Regardless of its shape, the sampling distribution of sample mean \bar{x} always has a mean identical to the sampled population, i.e. $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. This implies that *the spread of the distribution of sample means is considerably less than the spread of the sampled population*.

Central limit theorem A result that enables the use of normal probability distribution to approximate the sampling distribution of \bar{x} and \bar{p} .

The central limit theorem is useful in statistical inference. When the sample size is sufficiently large, estimations such as 'average' or 'proportion' that are used to make inferences about population parameters are expected to have sampling distribution that is approximately normal. The behaviour of these estimations can be described in repeated sampling and are used to evaluate the probability of observing certain sample results using the normal distribution as follows:

$$\text{Standard normal random variable, } z = \frac{\text{Estimator} - \text{Mean}}{\text{Standard deviation}}$$

As stated in the central limit theorem that the approximation to normal distribution is valid as long as the sample size is sufficiently 'large' – but how large? There is no clear understanding about the size of n . However, following guidelines are helpful in deciding an appropriate value of n :

- (i) If the sampled population is *normal*, then the sampling distribution of mean \bar{x} will also be normal, regardless of the size of sample.
- (ii) If the sampled population is approximately *symmetric*, then the sampling distribution of mean \bar{x} becomes approximately normal for relatively small values of n .
- (iii) If the sampled population is *skewed*, the sample size n must be larger, with at least 30 before the sampling distribution of mean \bar{x} becomes approximately normal

Thus the standard normal variate, $z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$ approximate the standard normal distribution, where μ and σ are the population mean and standard deviation, respectively.

7.8.2 Sampling Distribution of Mean When Population has Normal Distribution

Population Standard Deviation σ is Known As mentioned earlier that no matter what the population distribution is, for any given sample of size n taken from a population with mean μ and standard deviation σ , the sampling distribution of a sample statistic, such as mean and standard deviation are defined respectively by

- Mean of the distribution of sample means $\mu_{\bar{x}}$ or $E(\bar{x}) = \mu$ or expected value of the mean
- Standard deviation (or error) of the distribution of sample means $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ or standard error of the mean

If all possible samples of size n are drawn *with replacement* from a population having normal distribution with mean μ and standard deviation σ , then it can be shown that the sampling distribution of mean \bar{x} and standard error $\sigma_{\bar{x}}$ will also be normally distributed irrespective of the size of the sample. This result is true because any linear combination of normal random variables is also a normal random variable. In particular, if the sampling distribution of \bar{x} is normal, the standard error of the mean $\sigma_{\bar{x}}$ can be used in conjunction with normal distribution to determine the probabilities of various values of sample mean. For this purpose, the value of sample mean \bar{x} is first converted into a value z on the standard normal distribution to know how any single mean value deviates from the mean \bar{x} of sample mean values, by using the formula

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

since $\sigma_{\bar{x}}$ measures the dispersion (standard deviation) of values of sample means in the sampling distribution of the means, it can be said that

- $\bar{x} \pm \sigma_{\bar{x}}$ covers about the middle 68 per cent of the total possible sample means
- $\bar{x} \pm 1.96 \sigma_{\bar{x}}$ covers about the middle 95 per cent of the total possible sample means

The procedure for making statistical inference using sampling distribution about the population mean μ based on mean \bar{x} of sample means is summarized as follows:

- If the population standard deviation σ value is known and either
 - (a) population distribution is normal, or
 - (b) population distribution is not normal, but the sample size n is large ($n \geq 30$), then the sampling distribution of mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, is very close to the standard normal distribution given by

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- If the population is finite with N elements whose mean is μ and variance is σ^2 and the samples of fixed size n are drawn *without replacement*, then the standard deviation (also called standard error) of sampling distribution of mean \bar{x} can be modified to adjust the continued change in the size of the population N due to the several draws of samples of size n as follows:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Finite population correction factor The term $\sqrt{(N-n)/(N-1)}$ is multiplied with $\sigma_{\bar{x}}$ and $\sigma_{\bar{p}}$ a finite population is being sampled. In general, ignore the finite population correction factor whenever $n/N \leq 0.05$.

The term $\sqrt{(N-n)/(N-1)}$ is called the **finite population multiplier or finite correction factor**. In general, this factor has little effect on reducing the amount of sampling error when the size of the sample is less than 5 per cent of the population size. But if N is large relative to the sample size n , $\sqrt{(N-n)/(N-1)}$ is approximately equal to 1.

Population Standard Deviation σ is Not Known While calculating standard error $\sigma_{\bar{x}}$ of normally distributed sampling distribution, so far we have assumed that the population standard deviation σ is known. However, if σ is not known, the value of the normal variate z cannot be calculated for a specific sample. In such a case, the standard deviation of population σ must be estimated using the sample standard deviation s . Thus the standard error of the sampling distribution of mean \bar{x} becomes

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Since the value of $\sigma_{\bar{x}}$ varies according to each sample standard deviation, therefore instead of using the conversion formula

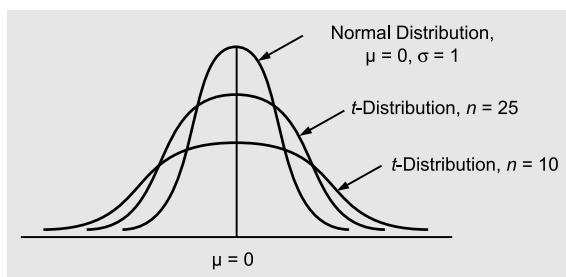
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

we use following formula, called 'Student's t -distribution'

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where $s = \sqrt{\sum(x - \bar{x})^2/(n-1)}$.

In contrast to the normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$, the t -distribution is a family of symmetrical distributions centred around the mean $\mu = 0$. The shape of the distribution depends on two statistics: \bar{x} and s . However, the value of s varies with sample size n . The higher the sample size n , s will be a more accurate estimate of population standard deviation σ and vice-versa. Figure 7.5. illustrates a comparison of t -distribution with that of the standard normal distribution



Degrees of freedom The number of unrestricted chances for variation in the measurement being made.

Degrees of Freedom The divisor $(n-1)$ in the formula for the sample variance s^2 is called number of *degrees of freedom (df)* associated with s^2 . The number of degrees of freedom refers to the *number of unrestricted chances for variation in the measurement being made*, i.e. number of independent squared deviations in s^2 that are available for estimating σ^2 . In other words, it refers to the number of values that are free to vary in a random sample. The shape of t -distribution varies with **degrees of freedom**. Obviously more is the sample size n , higher is the degrees of freedom.

Example 7.1: The mean length of life of a certain cutting tool is 41.5 hours with a standard deviation of 2.5 hours. What is the probability that a simple random sample of size 50 drawn from this population will have a mean between 40.5 hours and 42 hours?

[Delhi Univ., MBA, 2003]

Solution: We are given the following information

$$\mu = 41.5 \text{ hours}, \sigma = 2.5 \text{ hours}, \text{ and } n = 50$$

It is required to find the probability that the mean length of life, \bar{x} , of the cutting tool lies between 40.5 hours and 42 hours, that is, $P(40.5 \leq \bar{x} \leq 42)$.

Based upon the given information, the statistics of the sampling distribution are computed as:

$$\mu_{\bar{x}} = \mu = 41.5$$

and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{50}} = \frac{2.5}{7.0711} = 0.3536$

The population distribution is unknown, but sample size $n = 50$ is large enough to apply the central limit theorem. Hence, the normal distribution can be used to find the required probability as shown by the shaded area in Fig. 7.6.

$$\begin{aligned} P(40.5 \leq \bar{x} \leq 42) &= P\left[\frac{\bar{x}_1 - \mu}{\sigma_{\bar{x}}} \leq z \leq \frac{\bar{x}_2 - \mu}{\sigma_{\bar{x}}}\right] \\ &= P\left[\frac{40.5 - 41.5}{0.3536} \leq z \leq \frac{42 - 41.5}{0.3536}\right] \\ &= P[-2.8281 \leq z \leq 1.4140] \\ &= P[z \geq -2.8281] + P[z \leq 1.4140] \\ &= 0.4977 + 0.4207 = 0.9184 \end{aligned}$$

Thus 0.9184 is the probability of the tool of having a mean life between the required hours.

Example 7.2: A continuous manufacturing process produces items whose weights are normally distributed with a mean weight of 800 gms and a standard deviation of 300 gms. A random sample of 16 items is to be drawn from the process.

(a) What is the probability that the arithmetic mean of the sample exceeds 900 gms?
Interpret the results.

(b) Find the values of the sample arithmetic mean within which the middle 95 per cent of all sample means will fall.

Solution: (a) We are given the following information

$$\mu = 800 \text{ g}, \sigma = 300 \text{ g}, \text{ and } n = 16$$

Since population is normally distributed, the distribution of sample mean is normal with mean and standard deviation equal to

$$\mu_{\bar{x}} = \mu = 800$$

and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{300}{\sqrt{16}} = \frac{300}{4} = 75$

The required probability, $P(\bar{x} > 900)$ is represented by the shaded area in Fig. 7.7 of a normal curve. Hence

$$\begin{aligned} P(\bar{x} > 900) &= P\left[z > \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{900 - 800}{75}\right] \\ &= P[z > 1.33] \\ &= 0.5000 - 0.4082 = 0.0918 \end{aligned}$$

Hence, 6.18 per cent of all possible samples of size $n = 16$ will have a sample mean value greater than 900 g.

(b) Since $z = 1.96$ for the middle 95 per cent area under the normal curve as shown in Fig. 7.8, therefore using the formula for z to solve for the values of \bar{x} in terms of the known values are as follows:

$$\begin{aligned} \bar{x}_1 &= \mu_{\bar{x}} - z\sigma_{\bar{x}} \\ &= 800 - 1.96(75) = 653 \text{ g} \end{aligned}$$

and $\bar{x}_2 = \mu_{\bar{x}} + z\sigma_{\bar{x}}$
 $= 800 + 1.96(75) = 947 \text{ g}$

Figure 7.6
Normal curve

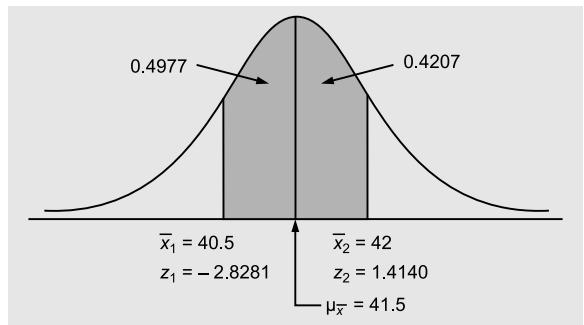


Figure 7.7
Normal curve

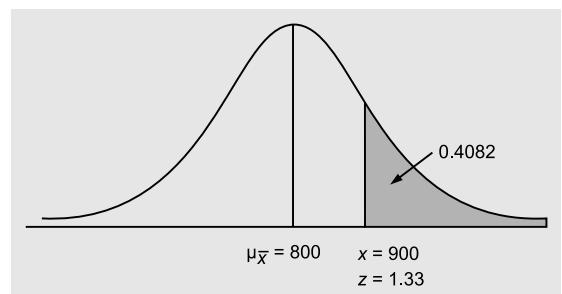
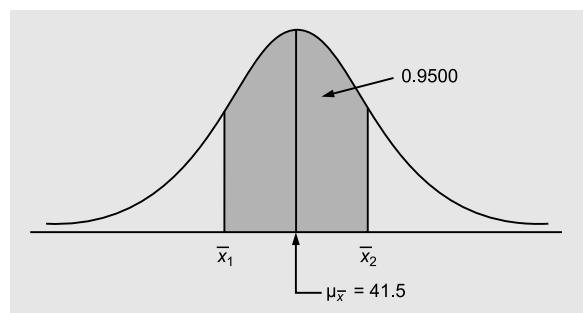


Figure 7.8
Normal curve



Example 7.3: An oil refinery has backup monitors to keep track of the refinery flows continuously and to prevent machine malfunctions from disrupting the process. One particular monitor has an average life of 4300 hours and a standard deviation of 730 hours. In addition to the primary monitor, the refinery has set up two standby units, which are duplicates of the primary one. In the case of malfunction of one of the monitors, another will automatically take over in its place. The operating life of each monitor is independent of the other.

- What is the probability that a given set of monitors will last at least 13,000 hours?
- At most 12,630 hours?

Solution: Given, $\mu = 4300$ hours, $\sigma = 730$ hours, $n = 3$. Based upon the given information the statistics of the sampling distribution are computed as:

$$\text{Mean, } \mu_{\bar{x}} = \mu = 4300$$

$$\text{and Standard deviation, } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{730}{\sqrt{3}} = \frac{730}{1.732} = 421.48$$

(a) For a set of monitors to last 13,000 hours, they must each last $13,000/3 = 4333.33$ hours on average. The required probability is calculated as follow:

$$\begin{aligned} P(\bar{x} \geq 4333.33) &= P\left[\frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \geq \frac{4333.33 - 4300}{421.48}\right] \\ &= P[z \geq 0.08] = 0.5 - 0.0319 = 0.4681 \end{aligned}$$

(b) For the set to last at most 12,630 hours, the average life can not exceed $12,630/3 = 4210$ hours. The required probability is calculated as follows:

$$\begin{aligned} P(\bar{x} \leq 4210) &= P\left[\frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \leq \frac{4210 - 4300}{421.48}\right] \\ &= P[z \leq -0.213] = 0.5 - 0.0832 = 0.4168 \end{aligned}$$

Example 7.4: Big Bazar, a chain of 130 shopping malls has been bought out by another larger nationwide supermarket chain. Before the deal is finalized, the larger chain wants to have some assurance that Big Bazar will be a consistent money maker. The larger chain has decided to look at the financial records of 25 of the Big Bazar outlets. Big Bazar claims that each outlet's profits have an approximately normal distribution with the same mean and a standard deviation of Rs 40 million. If the Big Bazar management is correct, then what is the probability that the sample mean for 25 outlets will fall within Rs 30 million of the actual mean?

Solution: Given $N = 130$, $n = 25$, $\sigma = 40$. Based upon the given information the statistics of the sampling distribution are computed as:

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{40}{\sqrt{25}} \sqrt{\frac{130-25}{130-1}} \\ &= \frac{40}{5} \sqrt{\frac{105}{129}} = 8 \times 0.902 = 13.72 \end{aligned}$$

The probability that the sample mean for 25 stores will fall within Rs 30 million is given by

$$\begin{aligned} P(\mu - 30 \leq \bar{x} \leq \mu + 30) &= P\left[\frac{-30}{13.72} \leq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \leq \frac{30}{13.72}\right] \\ &= P(-2.18 \leq z \leq 2.18) \\ &= 0.4854 + 0.4854 = 0.9708 \end{aligned}$$

Example 7.5: Chief Executive officer (CEO) of a life insurance company wants to undertake a survey of the huge number of insurance policies that the company has underwritten. The company makes an yearly profit on each policy that is distributed with mean Rs 8000 and standard deviation Rs 300. It is desired that the survey must be large enough to reduce the standard error to no more than 1.5 per cent of the population mean. How large should sample be?

Solution: Given $\mu = \text{Rs } 8000$, and $\sigma = \text{Rs } 300$. The aim is to find sample size n be large enough so that

$$\text{Standard error of estimate, } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \leq 1.5 \text{ per cent of Rs } 8000$$

$$\text{or } \frac{300}{\sqrt{n}} \leq 0.015 \times 8000 = 120$$

$$300 \leq 120\sqrt{n} \text{ or } \sqrt{n} \geq 25, \text{ or } n \geq 625$$

Thus, a sample size of at least 625 insurance policies is needed.

Example 7.6: Safal, a tea manufacturing company is interested in determining the consumption rate of tea per household in Delhi. The management believes that yearly consumption per household is normally distributed with an unknown mean μ and standard deviation of 1.50 kg

- (a) If a sample of 25 household is taken to record their consumption of tea for one year, what is the probability that the sample mean is within 500 gms of the population mean?
- (b) How large a sample must be in order to be 98 per cent certain that the sample mean is within 500 gms of the population mean?

Solution: Given $\mu = 500$ gms, $n = 25$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 1.5/\sqrt{25} = 0.25$ kg.

(a) Probability that the sample mean is within 500 gms or 0.5 kg of the population mean is calculated as follows:

$$\begin{aligned} P(\mu - 0.5 \leq \bar{x} \leq \mu + 0.5) &= P\left[\frac{-0.5}{\sigma/\sqrt{n}} \leq z \leq \frac{0.5}{\sigma/\sqrt{n}}\right] \\ &= P\left[\frac{-0.5}{0.25} \leq z \leq \frac{0.5}{0.25}\right] = P[-2 \leq z \leq 2] \\ &= 0.4772 + 0.4772 = 0.9544 \end{aligned}$$

(b) For 98 per cent confidence, the sample size is calculated as follows:

$$P(\mu - 0.5 \leq \bar{x} \leq \mu + 0.5) = P\left[\frac{-0.5}{1.5/\sqrt{n}} \leq z \leq \frac{0.5}{1.5/\sqrt{n}}\right]$$

Since $z = 2.33$ for 98 per cent area under normal curve, therefore

$$2.33 = \frac{0.5}{1.5/\sqrt{n}} \text{ or } 2.33 = 0.33\sqrt{n}$$

$$n = (2.33/0.33)^2 = 49.84$$

Hence, the management of the company should sample at least 50 households.

Example 7.7: A motorcycle manufacturing company claims that its particular brand of motorcycle gave an average highway km per litre rating of 90. An independent agency tested it to verify the claim. Under controlled conditions, the motorcycle was driven for a distance of 100 km on each of 25 different occasions. The actual kms per litre achieved during the trip were recorded on each occasion. Over the 25 trials, the average and standard deviation turned out to be 87 and 5 kms per litre, respectively. It is believed that the distribution of the actual highway km per litre for this motorcycle is close to a normal distribution.

If the rating of 90 km per litre of the agency is correct, find the probability that the average kms per litre over a random sample of 25 trials would be 87 or less.

Solution: Since the population standard deviation σ is unknown, t -Student's test will be applicable to calculate the desired probability $P(\bar{x} \leq 87)$ as follows:

$$\begin{aligned} P(\bar{x} \leq 87) &= P\left[t \leq \frac{\bar{x} - \mu}{s/\sqrt{n}}\right] = P\left[t \leq \frac{87 - 90}{5/\sqrt{25}}\right] \\ &= P[t \leq -3] \end{aligned}$$

with degrees of freedom $(n - 1) = (25 - 1) = 24$.

The desired probability of $t \leq -3.00$ with $df = 24$ from t -distribution table is 0.0031. Hence, the probability that the average km per litre is less than or equal to 87 is very small.

7.8.3 Sampling Distribution of Difference Between Two Sample Means

The concept of sampling distribution of sample mean introduced earlier in this chapter can also be used to compare a population of size N_1 having mean μ_1 and standard deviation σ_1 with another similar type of population of size N_2 having mean μ_2 and standard deviation σ_2 .

Let \bar{x}_1 and \bar{x}_2 be the mean of sampling distribution of mean of two populations, respectively. Then the difference between their mean values μ_1 and μ_2 can be estimated by generalizing the formula of standard normal variable as follows:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_{\bar{x}_1} - \mu_{\bar{x}_2})}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

where $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$ (mean of sampling distribution of difference of two means)

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (\text{Standard error of sampling distribution of two means})$$

n_1, n_2 = independent random samples drawn from first and second population, respectively.

Since random samples are drawn independently from two populations with replacement, therefore the sampling distribution of the difference of two means $\bar{x}_1 - \bar{x}_2$ will be normal provided sample size is sufficiently large.

The standard error of sampling distributions of some other statistic is given below:

Sampling Distribution	Standard Error and Mean	Remarks
• Median	$\sigma_{\text{Med}} = 1.2533 \frac{\sigma}{\sqrt{n}}$ $\mu_{\text{Med}} = \mu$	• For a large sample size $n \geq 30$, the sampling distribution of median approaches normal distribution. This result is true only if the population is normal or approximately normal.
• Sample standard deviation	(i) $\sigma_s = \frac{\sigma}{\sqrt{2n}}$ (ii) $\sigma_s = \sqrt{\frac{\mu_4 - \mu_2^2}{4n\mu_2}}$ (iii) $\mu_s = \sigma$	• For a large sample size $n \geq 100$, the sampling distribution is close to normal distribution. • If population is normally distributed, then σ_s is calculated using (i), otherwise (ii). • μ_2 and μ_4 are second and fourth moments, where $\mu_2 = \sigma^2$ and $\mu_4 = 3\sigma^4$.

Example 7.8: Car stereos of manufacturer A have a mean lifetime of 1400 hours with a standard deviation of 200 hours, while those of manufacturer B have a mean lifetime of 1200 hours with a standard deviation of 100 hours. If a random sample of 125 stereos of each manufacturer are tested, what is the probability that manufacturer A's stereos will have a mean lifetime which is at least (a) 160 hours more than manufacturer B's stereos and (b) 250 hours more than the manufacturer B's stereos? [Delhi Univ., MBA, 1999]

Solution: We are given the following information

Manufacturer A: $\mu_1 = 1400$ hours, $\sigma_1 = 200$ hours, $n_1 = 125$

Manufacturer B: $\mu_2 = 1200$ hours, $\sigma_2 = 100$ hours, $n_2 = 125$

Thus, $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2 = 1400 - 1200 = 200$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(200)^2}{125} + \frac{(100)^2}{125}} = \sqrt{80 + 320} = \sqrt{400} = 20$$

(a) Let $\bar{x}_1 - \bar{x}_2$ be the difference in mean lifetime of stereo manufactured by the two manufacturers. Then we are required to find the probability that this difference is more than or equal to 160 hours, as shown in Fig. 7.9. That is,

$$\begin{aligned} P[(\bar{x}_1 - \bar{x}_2) \geq 160] &= P\left[z \geq \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}\right] \\ &= P\left[z \geq \frac{160 - 200}{20}\right] \\ &= P[z \geq -2] \\ &= 0.5000 + 0.4772 = 0.9772 \text{ (Area under normal curve)} \end{aligned}$$

Hence, the probability is very high that the mean lifetime of the stereos of A is 160 hours more than that of B.

(b) Proceeding in the same manner as in part (a) as follows:

$$\begin{aligned} P[(\bar{x}_1 - \bar{x}_2) \geq 250] &= P\left[z \geq \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}\right] \\ &= P\left[z \geq \frac{250 - 200}{20}\right] = P[z \geq 2.5] \\ &= 0.5000 - 0.4938 \\ &= 0.0062 \\ &\quad \text{(Area under normal curve)} \end{aligned}$$

Hence, the probability is very less that the mean lifetime of the stereos of A is 250 hours more than that of B as shown in Fig. 7.10.

Example 7.9: The particular brand of ball bearings weighs 0.5 kg with a standard deviation of 0.02 kg. What is the probability that two lots of 1000 ball bearings each will differ in weight by more than 2 gms.

Solution: We are given the following information

Lot 1: $\mu_{\bar{x}_1} = \mu_1 = 0.50$ kg; $\sigma_1 = 0.02$ kg and $n_1 = 100$

Lot 2: $\mu_{\bar{x}_2} = \mu_2 = 0.50$ kg; $\sigma_2 = 0.02$ kg and $n_2 = 100$

Thus $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2 = 0$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(0.02)^2}{1000} + \frac{(0.02)^2}{1000}} = 0.000895$$

Figure 7.9
Normal curve

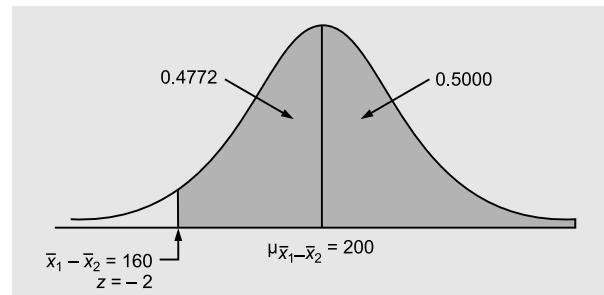


Figure 7.10
Normal curve

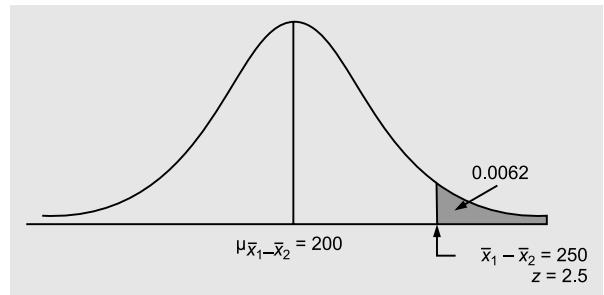
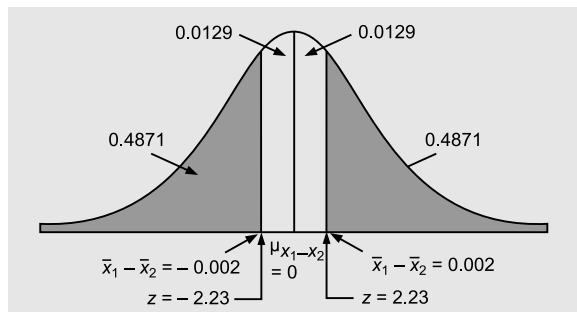


Figure 7.11
Normal curve



A difference of 2 gms in two lots is equivalent to a difference of $2/100 = 0.002$ kg in mean weights. It is possible if $\bar{x}_1 - \bar{x}_2 \leq 0.002$ or $\bar{x}_1 - \bar{x}_2 \geq -0.002$. Then the required probability that each ball bearing will differ by more than 2 gms is calculated as follows and shown in Fig. 7.11

$$P[-0.002 \leq \bar{x}_1 - \bar{x}_2 \leq 0.002]$$

$$\begin{aligned} &= P\left[\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} \leq z \leq \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}\right] \\ &= P\left[\frac{-0.002}{0.000895} \leq z \leq \frac{0.002}{0.000895}\right] \\ &= P[-2.33 \leq z \leq 2.33] \\ &= 2[0.5000 - 0.4871] = 0.0258 \end{aligned}$$

Self-Practice Problems 7A

- 7.1 A diameter of a component produced on a semi-automatic machine is known to be distributed normally with a mean of 10 mm and a standard deviation of 0.1 mm. If a random sample of size 5 is picked up, what is the probability that the sample mean will be between 9.95 mm and 10.05 mm?

[Delhi Univ., MBA, 1997]

- 7.2 The time between two arrivals at a queuing system is normally distributed with a mean of 2 minutes and standard deviation 0.25 minute. If a random sample of 36 is drawn, what is the probability that the sample mean will be greater than 2.1 minutes?

- 7.3 The strength of the wire produced by company A has a mean of 4,500 kg and a standard deviation of 200 kg. Company B has a mean of 4,000 kg and a standard deviation of 300 kg. If 50 wires of company A and 100 wires of company B are selected at random and tested for strength, what is the probability that the sample mean strength of A will be atleast 600 kg more than that of B? [Delhi Univ., MBA, 2000]

- 7.4 For a certain aptitude test, it is known from past experience that the average score is 1000 and the standard deviation is 125. If the test is administered to 100 randomly selected individuals, what is the probability that the value of the average score for this sample will lie in the interval 970 and 1030? Assume that the population distribution is normal.

- 7.5 A manufacturing process produces ball bearings with mean 5 cm and standard deviation 0.005 cm. A random sample of 9 bearings is selected to measure their average diameter and find it to be 5.004 cm. What is the probability that the average diameter of 9 randomly selected bearings would be at least 5.004 cm?

- 7.6 A population of items has an unknown distribution but a known mean and standard deviation of 50 and

100, respectively. Based upon a randomly drawn sample of 81 items drawn from the population, what is the probability that the sample arithmetic mean does not exceed 40?

- 7.7 A marketing research team has determined the standard error of sampling distribution of mean for a proposed market research sample size of 100 consumers. However, this standard error is twice the level that the management of the organization considers acceptable. What can be done to achieve an acceptable standard error for mean?

- 7.8 Assume that the height of 300 soldiers in an army batallion are normally distributed with mean 68 inches and standard deviation 3 inches. If 80 samples consisting of 25 soldiers each are taken, what would be the expected mean and standard deviation of the resulting sampling distribution of means if the sampling is done (a) with replacement and (b) without replacement?

- 7.9 How well have equity mutual funds performed in the past compared with BSE Stock Index? A random sample of 36 funds averages a 16.9 per cent annual investment return for 2001–2 with a standard deviation of 3.6 per cent annual return. The BSE Stock Index grew at an annual average rate of 16.3 per cent over the same period. Do these data show that, on the average, the mutual funds out-performed the BSE Stock Index during this period?

- 7.10 The average annual starting salary for an MBA is Rs 3,42,000. Assume that for the population of MBA (Marketing majors), the average annual starting salary is $\mu = 3,40,000$ and the standard deviation is $\sigma = 20,000$. What is the probability that a simple random sample of MBA (Marketing majors) will have a sample mean within \pm Rs 2,500 of the population mean for each sample sizes: 50,100 and 200? What is your conclusion? [Delhi Univ., MBA, 2003]

Hints and Answers

7.1 Given $\mu_{\bar{x}} = \mu = 10$, $\sigma = 0.1$ and $n = 10$. Thus

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.1/\sqrt{5} = 0.047$$

$$P[9.95 \leq \bar{x} \leq 10.05]$$

$$\begin{aligned} &= P\left[\frac{\bar{x}_1 - \mu}{\sigma_{\bar{x}_1}} \leq z \leq \frac{\bar{x}_2 - \mu}{\sigma_{\bar{x}_2}}\right] \\ &= P\left[\frac{9.95 - 10}{0.047} \leq z \leq \frac{10.05 - 10}{0.047}\right] \\ &= P[-1.12 \leq z \leq 1.12] \\ &= P[z \geq -1.12] + P[z \leq 1.12] \\ &= 0.3686 + 0.3686 = 0.7372 \end{aligned}$$

7.2 Given $\mu_{\bar{x}} = \mu = 2$, $\sigma = 0.25$ and $n = 36$. Thus

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} \frac{37,000 - 3,40,000}{20,000/\sqrt{50}} = 0.25/\sqrt{36} = 0.042$$

$$\begin{aligned} P[\bar{x} \geq 2.1] &= P\left[z \geq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\right] = P\left[z \geq \frac{2.1 - 2}{0.042}\right] \\ &= P[z \geq 2.38] = 0.5000 - 0.4913 \\ &= 0.0087 \end{aligned}$$

7.3 Given $\mu_1 = 4500$, $\sigma_1 = 200$ and $n_1 = 50$; $\mu_2 = 4000$, $\sigma_2 = 300$ and $n_2 = 100$. Then

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2 = 4500 - 4000 = 500$$

$$\begin{aligned} \sigma_{\bar{x}_1 - \bar{x}_2} &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{40,000}{50} + \frac{90,000}{100}} \\ &= 41.23 \end{aligned}$$

$$\begin{aligned} P[(\bar{x}_1 - \bar{x}_2) \geq 600] &= P\left[z \geq \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}\right] \\ &= P\left[z \geq \frac{600 - 500}{41.23}\right] \\ &= P(z \geq 2.43) \\ &= 0.5000 - 0.4925 = 0.0075 \end{aligned}$$

7.4 Given $\mu_{\bar{x}} = \mu = 1000$, $\sigma = 125$ and $n = 100$. Thus $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 125/\sqrt{100} = 12.5$

$$P(970 \leq \bar{x} \leq 1030)$$

$$\begin{aligned} &= P\left[\frac{\bar{x}_1 - \mu}{\sigma_{\bar{x}}} \leq z \leq \frac{\bar{x}_2 - \mu}{\sigma_{\bar{x}}}\right] \\ &= P\left[\frac{970 - 1000}{12.5} \leq z \leq \frac{1030 - 1000}{12.5}\right] \\ &= P(-2.4 \leq z \leq 2.4) \\ &= P(z \leq 2.4) + P(z \geq -2.4) \\ &= 0.4918 + 0.4918 = 0.9836 \end{aligned}$$

7.5 Given $\mu_{\bar{x}} = \mu = 5$, $\sigma = 0.005$ and $n = 9$. Thus

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.005/\sqrt{9} = 0.0017$$

$$\begin{aligned} P(\bar{x} \geq 5.004) &= P\left[z \geq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\right] \\ &= P\left[z \geq \frac{5.004 - 5.000}{0.0017}\right] \\ &= P(z \geq 2.4) = 1 - P(z \leq 2.4) \\ &= 1 - 0.9918 = 0.0082 \end{aligned}$$

7.6 Given $\mu_{\bar{x}} = \mu = 50$, $\sigma = 100$ and $n = 81$. Thus

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 100/\sqrt{81} = 11.1$$

$$\begin{aligned} P(\bar{x} \leq 40) &= P\left[z \geq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\right] = P\left[z \leq \frac{40 - 50}{11.1}\right] \\ &= P(z \leq -0.90) = 0.5000 - 0.3159 = 0.1841 \end{aligned}$$

7.7 Since standard error is inversely proportional to the square root of the sample size, therefore to reduce the standard error determined by the market research team, the sample size should be increased to $n = 400$ (four times of $n = 100$).

7.8 The number of possible samples of size 25 each from a group of 3000 soldiers with and without replacement are $(3000)^{25}$ and ${}^{300}C_{25}$, respectively. These numbers are much larger than 80—actually drawn samples. Thus we will get only an experimental sampling distribution of means rather than true sampling distribution. Hence mean and standard deviation would be close to those of the theoretical distribution. That is:

$$(a) \mu_{\bar{x}} = \mu = 68 \text{ and } \sigma_{\bar{x}} = \sigma/\sqrt{n} = 3/\sqrt{25} = 0.60$$

$$\begin{aligned} (b) \mu_{\bar{x}} &= \mu = 68 \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \\ &= \frac{3}{\sqrt{25}} \sqrt{\frac{3000-25}{3000-1}} = 1.19 \end{aligned}$$

7.9 Given $\mu_{\bar{x}} = \mu = 16.9$, $\sigma = 3.6$ and $n = 36$. Thus

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 3.6/\sqrt{36} = 0.60$$

$$\begin{aligned} P(\bar{x} \geq 16.3) &= P\left[z \geq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\right] = P\left[z \geq \frac{16.3 - 16.9}{0.60}\right] \\ &= P[z \geq -1] = 0.5000 + 0.1587 = 0.6587 \end{aligned}$$

7.10 Given $\mu = 3,40,000$; $\sigma = 20,000$, $n_1 = 50$, $n_2 = 100$, and $n_3 = 200$

For $n_1 = 50$:

$$z_1 = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{3,42,000 - 3,40,000}{20,000/\sqrt{50}} = 0.88$$

$$z_2 = \frac{3,37,000 - 3,40,000}{20,000/\sqrt{50}} = -0.88$$

$$P(-0.88 \leq z \leq 0.88) = 0.3106 \times 2 = 0.6212$$

Similar calculations for $n_2 = 100$ and $n_2 = 200$ give

$$P(-1.25 \leq z \leq 1.25) = 0.3944 \times 2 = 0.7888$$

$$P(-1.76 \leq z \leq 1.76) = 0.4616 \times 2 = 0.9282$$

7.9 SAMPLING DISTRIBUTION OF SAMPLE PROPORTION

There are many situations in which each element of the population can be classified into two mutually exclusive categories such as success or failure, accept or reject, head or tail of a coin, and so on. These and similar situations provide practical examples of binomial experiments, if the sampling procedure has been conducted in an appropriate manner. If a random sample of n elements is selected from the binomial population and x of these possess the specified characteristic, then the sample proportion \bar{p} is the best statistic to use for statistical inferences about the population proportion parameter p . The sample proportion can be defined as:

$$\bar{p} = \frac{\text{Elements of sample having characteristic, } x}{\text{Sample size, } n}$$

With the same logic of sampling distribution of mean, the sampling distribution of sample proportions with mean $\mu_{\bar{p}}$ and standard deviation (also called *standard error*) $\sigma_{\bar{p}}$ is given by

$$\mu_{\bar{p}} = p \text{ and } \sigma_{\bar{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$$

If the sample size n is large ($n \geq 30$), the sampling distribution of \bar{p} can be approximated by a normal distribution. The approximation will be adequate if

$$np \geq 5 \text{ and } n(1-p) \geq 5$$

It may be noted that the sampling distribution of the proportion would actually follow binomial distribution because population is binomially distributed.

The mean and standard deviation (error) of the sampling distribution of proportion are valid for a finite population in which sampling is with replacement. However, for finite population in which sampling is done without replacement, we have

$$\mu_{\bar{p}} = p \text{ and } \sigma_{\bar{p}} = \sqrt{\frac{pq}{n}} \sqrt{\frac{N-n}{N-1}}$$

Under the same guidelines as mentioned in previous sections, for a large sample size n (≥ 30), the sampling distribution of proportion is closely approximated by a normal distribution with mean and standard deviation as stated above. Hence, to standardize sample proportion \bar{p} , the standard normal variable.

$$z = \frac{\bar{p} - \mu_{\bar{p}}}{\sigma_{\bar{p}}} = \frac{\bar{p} - p}{\sqrt{p(1-p)/n}}$$

is approximately the standard normal distribution.

7.9.1 Sampling Distribution of the Difference of Two Proportions

Suppose two populations of size N_1 and N_2 are given. For each sample of size n_1 from first population, compute sample proportion \bar{p}_1 and standard deviation $\sigma_{\bar{p}_1}$. Similarly, for each sample of size n_2 from second population, compute sample proportion \bar{p}_2 and standard deviation $\sigma_{\bar{p}_2}$.

For all combinations of these samples from these populations, we can obtain a sampling distribution of the difference $\bar{p}_1 - \bar{p}_2$ of samples proportions. Such a distribution is called *sampling distribution of difference of two proportions*. The mean and standard deviation of this distribution are given by

$$\mu_{\bar{p}_1 - \bar{p}_2} = \mu_{\bar{p}_1} - \mu_{\bar{p}_2} = p_1 - p_2$$

$$\text{and } \sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\sigma_{\bar{p}_1}^2 + \sigma_{\bar{p}_2}^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

If sample size n_1 and n_2 are large, that is, $n_1 \geq 30$ and $n_2 \geq 30$, then the sampling distribution of difference of proportions is closely approximated by a normal distribution.

Example 7.10: A manufacturer of watches has determined from experience that 3 per cent of the watches he produces are defective. If a random sample of 300 watches is examined, what is the probability that the proportion defective is between 0.02 and 0.035? [Delhi Univ., MBA, 1990]

Solution: We are given the following information

$$\mu_{\bar{p}} = p = 0.03, \bar{p}_1 = 0.02, \bar{p}_2 = 0.035 \text{ and } n = 300$$

Thus standard error of proportion is given by

$$\begin{aligned}\sigma_{\bar{p}} &= \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.03 \times 0.97}{300}} \\ &= \sqrt{0.000097} = 0.0098\end{aligned}$$

For calculating the desired probability, we apply the following formula

$$\begin{aligned}P[0.02 \leq \bar{p} \leq 0.035] &= P\left[\frac{\bar{p}_1 - p}{\sigma_{\bar{p}}} \leq z \leq \frac{\bar{p}_2 - p}{\sigma_{\bar{p}}}\right] \\ &= P\left[\frac{0.02 - 0.03}{0.0098} \leq z \leq \frac{0.035 - 0.03}{0.0098}\right] \\ &= P[-1.02 \leq z \leq 0.51] \\ &= P(z \geq -1.02) + P(z \leq 0.51) = 0.3461 + 0.1950 = 0.5411\end{aligned}$$

Hence the probability that the proportion of defectives will lie between 0.02 and 0.035 is 0.5411.

Example 7.11: Few years back, a policy was introduced to give loan to unemployed engineers to start their own business. Out of 1,00,000 unemployed engineers, 60,000 accept the policy and got the loan. A sample of 100 unemployed engineers is taken at the time of allotment of loan. What is the probability that sample proportion would have exceeded 50 per cent acceptance?

Solution: We are given the following information

$$\mu_{\bar{p}} = p = 0.60, N = 1,00,000 \text{ and } n = 100$$

Thus the standard error of proportion in a finite population of size 1,00,000 is given by

$$\begin{aligned}\sigma_{\bar{p}} &= \sqrt{\frac{p(1-p)}{n} \sqrt{\frac{N-n}{N-1}}} = \sqrt{\frac{0.60 \times 0.40}{100} \sqrt{\frac{1,00,000 - 100}{1,00,000 - 1}}} \\ &= \sqrt{0.0024} \sqrt{0.9990} = 0.0489 \times 0.9995 = 0.0488\end{aligned}$$

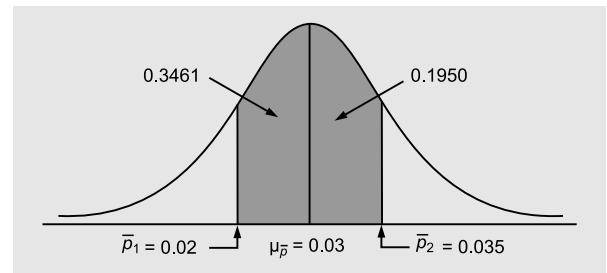
The probability that sample proportion would have exceeded 50 per cent acceptance is given by

$$\begin{aligned}P(x \geq 0.50) &= P\left[z \geq \frac{\bar{p} - p}{\sigma_{\bar{p}}}\right] = P\left[z \geq \frac{0.50 - 0.60}{0.0489}\right] \\ &= P[z \geq -2.04] = 0.5000 + 0.4793 = 0.9793\end{aligned}$$

Example 7.12: Ten per cent of machines produced by company A are defective and five per cent of those produced by company B are defective. A random sample of 250 machines is taken from company A and a random sample of 300 machines from company B. What is the probability that the difference in sample proportion is less than or equal to 0.02?

[South Gujarat Univ, MBA; Delhi Univ., MBA, 1999]

Figure 7.12
Normal curve



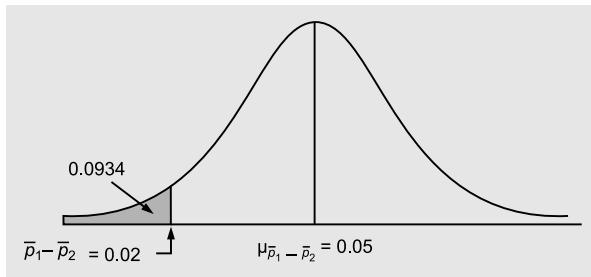
Solution: We are given the following information

$$\mu_{\bar{p}_1 - \bar{p}_2} = \mu_{\bar{p}_1} - \mu_{\bar{p}_2} = p_1 - p_2 = 0.10 - 0.05 = 0.05; n_1 = 250 \text{ and } n_2 = 300$$

Thus standard error of the difference in a sample proportion is given by

$$\begin{aligned}\mu_{\bar{p}_1 - \bar{p}_2} &= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{0.10 \times 0.90}{250} + \frac{0.05 \times 0.95}{300}} \\ &= \sqrt{\frac{0.90}{250} + \frac{0.0475}{300}} = \sqrt{0.00052} = 0.0228\end{aligned}$$

Figure 7.13
Normal curve



The desired probability of difference in sample proportions is given by

$$\begin{aligned}P[(\bar{p}_1 - \bar{p}_2) \leq 0.02] &= P\left[z \leq \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sigma_{\bar{p}_1 - \bar{p}_2}}\right] \\ &= P\left[z \leq \frac{0.02 - 0.05}{0.0228}\right] = P[z \leq -1.32] \\ &= 0.5000 - 0.4066 = 0.0934\end{aligned}$$

Hence the desired probability for the difference in sample proportions is 0.0934.

Self-Practice Problems 7B

- 7.11** Assume that 2 per cent of the items produced in an assembly line operation are defective, but that the firm's production manager is not aware of this situation. What is the probability that in a lot of 400 such items, 3 per cent or more will be defective?
- 7.12** If a coin is tossed 20 times and the coin falls on head after any toss, it is a success. Suppose the probability of success is 0.5. What is the probability that the number of successes is less than or equal to 12?
- 7.13** The quality control department of a paints manufacturing company, at the time of despatch of decorative paints, discovered that 30 per cent of the containers are defective. If a random sample of 500 containers is drawn with replacement from the population, what is the probability that the sample

proportion will be less than or equal to 25 per cent defective?

- 7.14** A manufacturer of screws has found that on an average 0.04 of the screws produced are defective. A random sample of 400 screws is examined for the proportion of defective screws. Find the probability that the proportion of defective screws in the sample is between 0.02 and 0.05.
- 7.15** A manager in the billing section of a mobile phone company checks on the proportion of customers who are paying their bills late. Company policy dictates that this proportion should not exceed 20 per cent. Suppose that the proportion of all invoices that were paid late is 20 per cent. In a random sample of 140 invoices, determine the probability that more than 28 per cent invoices were paid late.

Hints and Answers

7.11 $\mu_{\bar{p}} = np = 400 \times 0.02 = 8;$

$$\sigma_{\bar{p}} = \sqrt{npq} = \sqrt{400 \times 0.02 \times 0.98} = 2.8$$

and 3% of 400 = 12 defective items. Thus

$$\begin{aligned}P(\bar{p} \geq 12) &= P\left[z \geq \frac{\bar{p} - np}{\sigma_{\bar{p}}}\right] = P\left[z \geq \frac{12 - 8}{2.8}\right] \\ &= P(z \geq 1.42) = 0.5000 - 0.4222 \\ &= 0.0778\end{aligned}$$

7.12 Given $\mu_{\bar{p}} = np = 20 \times 0.50 = 10;$

$$\sigma_{\bar{p}} = \sqrt{npq} = \sqrt{20 \times 0.50 \times 0.50} = 2.24$$

$$\begin{aligned}P(\bar{p} \leq 12) &= P\left[z \leq \frac{\bar{p} - np}{\sigma_{\bar{p}}}\right] = P\left[z \leq \frac{12 - 10}{2.24}\right] \\ &= P(z \leq 0.89) = 0.8133\end{aligned}$$

- 7.13 Given $\mu_{\bar{p}} = p = 0.30, n = 500$;

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.30 \times 0.70}{500}} = 0.0205.$$

$$\begin{aligned} P(\bar{p} \leq 0.25) &= P\left[z \leq \frac{\bar{p} - p}{\sigma_{\bar{p}}}\right] = \left[z \leq \frac{0.25 - 0.30}{0.0205}\right] \\ &= P[z \leq -2.43] = 0.5000 - 0.4927 \\ &= 0.0083 \end{aligned}$$

- 7.14 Given $\mu_{\bar{p}} = p = 0.04, n = 400$;

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.04 \times 0.96}{400}} = 0.009$$

$$P[0.02 \leq \bar{p} \leq 0.05] = P\left[\frac{\bar{p}_1 - p}{\sigma_{\bar{p}}} \leq z \leq \frac{\bar{p}_2 - p}{\sigma_{\bar{p}}}\right]$$

$$\begin{aligned} &= P\left[\frac{0.02 - 0.04}{0.009} \leq z \leq \frac{0.05 - 0.04}{0.009}\right] \\ &= P[-2.22 \leq z \leq 2.22] \\ &= P[z \geq -2.22] + P[z \leq 2.22] \\ &= 0.4861 + 0.4861 = 0.9722 \end{aligned}$$

- 7.15 Given $\mu_{\bar{p}} = p = 0.20, n = 140$;

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.20 \times 0.80}{140}} = 0.033$$

$$\begin{aligned} P[\bar{p} \geq 0.28] &= P\left[z \geq \frac{\bar{p} - p}{\sigma_{\bar{p}}}\right] \\ &= P\left[z \geq \frac{0.28 - 0.20}{0.033}\right] = P[z \geq 2.42] \\ &= 0.5000 - 0.4918 = 0.0082 \end{aligned}$$

Formulae Used

1. Standard deviation (or standard error) of sampling distribution of mean, \bar{x}

- Infinite Population: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

- Finite Population: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

where $n < 0.5N$; n, N = size of sample and population, respectively.

2. Estimate of $\sigma_{\bar{x}}$ when population standard deviation is not known

- Infinite Population: $s_{\bar{x}} = \frac{s}{\sqrt{n}}$

- Finite Population: $s_{\bar{x}} = \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

3. Standard deviation of sampling distribution of sample means

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

4. Standard deviation (or standard error) of sampling distribution of proportion

- Infinite Population: $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} ; q = 1-p$

- Finite Population: $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$

5. Standard deviation of sampling distribution of sample proportions

$$\sigma_{p_1 - p_2} = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}};$$

$$q_1 = 1 - p_1; q_2 = 1 - p_2$$

Review Self-Practice Problems

- 7.16 An auditor takes a random sample of size $n = 36$ from a population of 1000 accounts receivable. The mean value of the accounts receivable for the population is Rs 260 with a population standard deviation of Rs 45. What is the probability that the sample mean will be less than Rs 250?

- 7.17 A marketing research analyst selects a random sample of 100 customers out of the 400 who purchased a particular item from central store. The 100 customers spent an average of Rs 250 with a standard deviation

of Rs 70. For a middle 95 per cent customers, determine the mean purchase amount for all 400 customers.

- 7.18 In a particular coal mine, 5000 employees on an average are of 58 years of age with a standard deviation of 8 years. If a random sample of 50 employees is taken, what is the probability that the sample will have an average age of less than 60 years?

- 7.19 A simple random sample of 50 ball bearings taken from a large number being manufactured has a mean weight of 1.5 kg per bearing with a standard deviation of 0.1 kg.

- (a) Estimate the value of the standard error of the mean
 (b) If the sample of 50 ball bearings is taken from a particular production run that includes just 150 bearings as the total population, then estimate the standard error of the mean and compare it with the result of part (a).
- 7.20** A population proportion is 0.40. A simple random sample of size 200 will be taken and the sample proportion will be used to estimate the population proportion, what is the probability that the sample proportion will be within ± 0.03 of the population proportion?
- 7.21** A sales manager of a firm believes that 30 per cent of the firm's orders come from first time customers. A simple random sample of 100 orders will be used to estimate the proportion of first-time customers. Assume that the sales manager is correct and proportion is 0.30.
- (a) Justify sampling distribution of proportion for this case
 (b) What is probability that the sample proportion will be between 0.20 and 0.40?
- 7.22** The diameter of a steel pipe manufactured at a large factory is expected to be approximately normally distributed with a mean of 1.30 inches and a standard deviation of 0.04 inch.
 (a) If a random sample of 16 pipes is selected, then what is the probability that randomly selected pipe will have a diameter between 1.28 and 1.30 inches?
 (b) Between what two values will 60 per cent of the pipes fall in terms of the diameter?

Hints and Answers

7.16 Given $\mu_{\bar{x}} = \mu = 260$; $n = 36$;

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 45/\sqrt{36} = 7.5$$

$$\begin{aligned} P(\bar{x} \leq 250) &= P\left[z \leq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\right] = P\left[z \leq \frac{250 - 260}{7.5}\right] \\ &= P(z \leq -1.33) \\ &= 0.5000 - 0.4082 = 0.0918 \end{aligned}$$

7.17 Given $s = 70$, $n = 100$, $\bar{x} = 250$, $z = 1.96$ at 95% confidence. Thus

$$\begin{aligned} s_{\bar{x}} &= \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{70}{\sqrt{100}} \sqrt{\frac{400-100}{400-1}} \\ &= 7(0.867) = 11.33 \\ \bar{x} \pm z s_{\bar{x}} &= 250 \pm 1.96(11.33) \\ &= \text{Rs } 227.80 \text{ to Rs } 272.20 \end{aligned}$$

7.18 Given $n = 50$, $N = 5000$, $\mu = 58$, and $\sigma = 8$

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{8}{\sqrt{50}} \sqrt{\frac{5000-50}{5000-1}} \\ &= 1.131 \times 0.995 = 1.125 \\ P(\bar{x} \leq 60) &= P\left[z \leq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\right] = P\left[z \leq \frac{58-60}{1.125}\right] \\ &= P(z \leq -1.77) \\ &= 0.5000 - 0.4616 = 0.0384 \end{aligned}$$

7.19 Given $\mu_{\bar{x}} = \mu = 1.5$, $n = 50$, $N = 150$ and $s = 0.1$.

$$(a) s_{\bar{x}} = s/\sqrt{n} = 0.1/\sqrt{50} = 0.014 \text{ kg}$$

$$(b) s_{\bar{x}} = \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = 0.014 \sqrt{\frac{150-50}{150-1}} = 0.011 \text{ kg.}$$

It is less than the value in part (a) due to finite correction factor.

7.20 Given $\mu_{\bar{p}} = \bar{p} = 0.40$, $n = 200$

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.40 \times 0.60}{200}} = 0.0346$$

(a) Justify sampling distribution of proportion for this case

(b) What is probability that the sample proportion will be between 0.20 and 0.40?

7.22 The diameter of a steel pipe manufactured at a large factory is expected to be approximately normally distributed with a mean of 1.30 inches and a standard deviation of 0.04 inch.

(a) If a random sample of 16 pipes is selected, then what is the probability that randomly selected pipe will have a diameter between 1.28 and 1.30 inches?

(b) Between what two values will 60 per cent of the pipes fall in terms of the diameter?

$$\begin{aligned} P(-0.03 \leq \bar{p} \leq 0.03) &= 2P\left[z \leq \frac{\bar{p} - p}{\sigma_{\bar{p}}}\right] \\ &= 2P\left[z \leq \frac{0.03}{0.0346}\right] \\ &= 2P(z \leq 0.87) \\ &= 2 \times 0.3078 = 0.6156 \end{aligned}$$

7.21 Given $p = 0.30$, $n = 100$. Thus

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.30 \times 0.70}{100}} = 0.0458$$

(a) Since both $np = 100(0.30) = 30$ and $nq = n(1-p) = 100(0.70) = 70$ are greater than 5, the normal distribution is appropriate to use

(b) $P(0.20 \leq \bar{p} \leq 0.40)$

$$\begin{aligned} &= P\left[\frac{0.20 - 0.30}{0.0458} \leq z \leq \frac{0.40 - 0.30}{0.0458}\right] \\ &= P[-2.18 \leq z \leq 2.18] \\ &= 2P(z \leq 2.18) = 2 \times 0.4854 = 0.9708 \end{aligned}$$

7.22 Given $\mu_{\bar{x}} = \mu = 1.30$, $\sigma = 0.04$ and $n = 16$,

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.04/\sqrt{16} = 0.01$$

$$(a) P(1.28 \leq \bar{x} \leq 1.30) = P\left[\frac{\bar{x}_1 - \mu}{\sigma_{\bar{x}}} \leq z \leq \frac{\bar{x}_2 - \mu}{\sigma_{\bar{x}}}\right]$$

$$= P\left[\frac{1.28 - 1.30}{0.01} \leq z \leq \frac{1.30 - 1.30}{0.01}\right]$$

$$= P[2 \leq z \leq 0] = 0.5000 - 0.4772$$

$$= 0.0228$$

$$\begin{aligned} (b) \bar{x} \pm z \sigma_{\bar{x}} &= 1.30 \pm 0.84(0.01) = 1.30 \pm 0.0084 \\ &= 1.2916 \text{ to } 1.3084 \end{aligned}$$

If there is an opportunity to make a mistake, sooner or later the mistake will be made.

—Edmond C Berekely

Hypothesis Testing

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- explain why hypothesis testing is important
- know how to establish null and alternative hypotheses about a population parameter
- develop hypothesis testing methodology for accepting or rejecting null hypothesis.
- use the test statistic z , t , and F to test the validity of a claim or assertion about the true value of any population parameter.
- understand Type I and Type II errors and its implications in making a decision.
- compute and interpret p-values
- interpret the confidence level, the significance level, and the power of a test

8.1 INTRODUCTION

Inferential statistics is concerned with estimating the true value of population parameters using sample statistics. Three techniques of inferential statistics, namely, a (i) *point estimate*, (ii) *confidence interval that is likely to contain the true parameter value*, and (iii) *degree of confidence associated with a parameter value which lies within an interval values*. This information helps decision-makers in determining an interval estimate of a population parameter value with the help of sample statistic along with certain level of confidence of the interval containing the parameter value. Such an estimate is helpful for drawing statistical inference about the characteristic (or feature) of the population of interest.

Another way of estimating the true value of population parameters is to test the validity of the claim (assertion or statement) made about this true value using sample statistics.

8.2 HYPOTHESIS AND HYPOTHESIS TESTING

A *statistical hypothesis* is a claim (assertion, statement, belief or assumption) about an unknown population parameter value. For example (i) a judge assumes that a person

charged with a crime is innocent and subject this assumption (hypothesis) to a verification by reviewing the evidence and hearing testimony before reaching to a verdict (ii) a pharmaceutical company claims the efficacy of a medicine against a disease that 95 per cent of all persons suffering from the said disease get cured (iii) an investment company claims that the average return across all its investments is 20 per cent, and so on. To test such claims or assertions statistically, sample data are collected and analysed. On the basis of sample findings the hypothesized value of the population parameter is either accepted or rejected. *The process that enables a decision maker to test the validity (or significance) of his claim by analysing the difference between the value of sample statistic and the corresponding hypothesized population parameter value, is called hypothesis testing.*

8.2.1 Formats of Hypothesis

As stated earlier, a hypothesis is a statement to be tested about the true value of population parameter using sample statistics. A hypothesis whether there exists any significant difference between two or more populations with respect to any of their common parameter can also be tested. To examine whether any difference exists or not, a hypothesis can be stated in the form of *if-then* statement. Consider, for instance, the nature of following statements:

- If inflation rate has decreased, then wholesale price index will also decrease.
- If employees are healthy, then they will take sick leave less frequently.

If terms such as ‘positive,’ ‘negative,’ ‘more than,’ ‘less than,’ etc are used to make a statement, then such a hypothesis is called *directional hypothesis* because it indicates the direction of the relationship between two or more populations under study with respect to a parameter value as illustrated below:

- Side effects were experienced by less than 20 per cent of people who take a particular medicine.
- Greater the stress experienced in the job, lower the job satisfaction to employees.

The *nondirectional hypothesis* indicates a relationship (or difference), but offer no indication of the direction of relationships (or differences). In other words, though it may be obvious that there would be a significant relationship between two populations with respect to a parameter, we may not be able to say whether the relationship would be positive or negative. Similarly, even if we consider that two populations differ with respect to a parameter, it will not be easy to say which population will be more or less. Following examples illustrate non-directional hypotheses.

- There is a relationship between age and job satisfaction.
- There is a difference between average pulse rates of men and women.

8.3 THE RATIONALE FOR HYPOTHESIS TESTING

The inferential statistics is concerned with estimating the unknown population parameter by using sample statistics. If a claim or assumption is made about the specific value of population parameter, then it is expected that the corresponding sample statistic is close to the hypothesized parameter value. It is possible only if hypothesized parameter value is correct and the sample statistic turns out to be a good estimator of the parameter. This approach to test a hypothesis is called a *test statistic*.

Since sample statistics are random variables, therefore their sampling distributions show the tendency of variation. Consequently we do not expect the sample statistic value to be equal to the hypothesized parameter value. The difference, if any, is due to chance and/or sampling error. But if the value of the sample statistic differs significantly from the hypothesized parameter value, then the question arises whether the hypothesized parameter value is correct or not. The greater the difference between the value of the sample statistic and hypothesized parameter, the more doubt is there about the correctness of the hypothesis.

In statistical analysis, difference between the value of the sample statistic and hypothesized parameter is specified in terms of the given level of probability whether

the particular level of difference is significant or not when the hypothesized parameter value is correct. The probability that a particular level of deviation occurs by chance can be calculated from the known sampling distribution of the test statistic.

The probability level at which the decision-maker concludes that observed difference between the value of the test statistic and hypothesized parameter value cannot be due to chance is called the *level of significance* of the test.

8.4 GENERAL PROCEDURE FOR HYPOTHESIS TESTING

As mentioned before, to test the validity of the claim or assumption about the population parameter, a sample is drawn from the population and analysed. The results of the analysis are used to decide whether the claim is true or not. The steps of general procedure for any hypothesis testing are summarized below:

Step 1: State the Null Hypothesis (H_0) and Alternative Hypothesis (H_1)

The **null hypothesis** H_0 (read as H_0 sub-zero) represents the claim or statement made about the value or range of values of the population parameter. The capital letter H stands for hypothesis and the subscript 'zero' implies 'no difference' between sample statistic and the parameter value. Thus hypothesis testing requires that the null hypothesis be considered *true (status quo or no difference)* until it is proved false on the basis of results observed from the sample data. The null hypothesis is always expressed in the form of mathematical statement which includes the sign (\leq , $=$, \geq) making a claim regarding the specific value of the population parameter. That is:

$$H_0 : \mu (\leq, =, \geq) \mu_0$$

where μ is population mean and μ_0 represents a hypothesized value of μ . Only one sign out of \leq , $=$ and \geq will appear at a time when stating the null hypothesis

An **alternative hypothesis**, H_1 , is the counter claim (statement) made against the value of the particular population parameter. That is, an alternative hypothesis must be true when the null hypothesis is found to be false. In other words, the alternative hypothesis states that specific population parameter value is not equal to the value stated in the null hypothesis and is written as:

$$H_1 : \mu \neq \mu_0$$

Consequently $H_1 : \mu < \mu_0$ or $H_1 : \mu > \mu_0$

Each of the following statements is an example of a null hypothesis and alternative hypothesis:

- | |
|--|
| <ul style="list-style-type: none"> • $H_0 : \mu = \mu_0$; $H_1 : \mu \neq \mu_0$ • $H_0 : \mu \leq \mu_0$; $H_1 : \mu > \mu_0$ • $H_0 : \mu \geq \mu_0$; $H_1 : \mu < \mu_0$ |
|--|

(a) Directional hypothesis

- H_0 : There is no difference between the average pulse rates of men and women
 H_1 : Men have lower average pulse rates than women do
- H_0 : There is no relationship between exercise intensity and the resulting aerobic benefit
 H_1 : Increasing exercise intensity increases the resulting aerobic benefit
- H_0 : The defendant is innocent
 H_1 : The defendant is guilty

(b) Non-Directional hypothesis

- H_0 : Men and Women have same verbal abilities
 H_1 : Men and women have different verbal abilities
- H_0 : The average monthly salary for management graduates with a 4-year experience is Rs 75,000.
 H_1 : The average monthly salary is not Rs 75,000.

Null hypothesis: The hypothesis which is initially assumed to be true, although it may in fact be either true or false based on the sample data.

Alternative hypothesis: The hypothesis concluded to be true if the null hypothesis is rejected.

- H_0 : Older workers are more loyal to a company
- H_1 : Older workers may not be loyal to a company

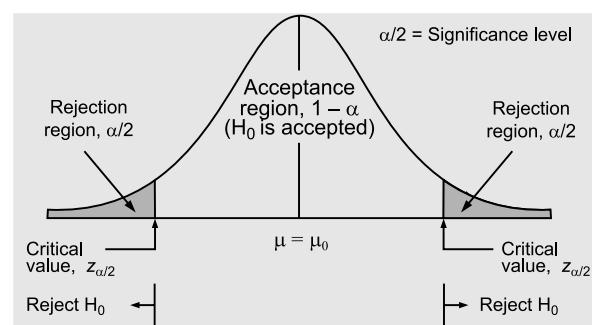
Step 2: State the Level of Significance, α (alpha)

The level of significance, usually denoted by α (alpha), is specified before the samples are drawn, so that the results obtained should not influence the choice of the decision-maker. It is specified in terms of the probability of null hypothesis H_0 being wrong. In other words, the level of significance defines the likelihood of rejecting a null hypothesis when it is true, i.e. it is *the risk a decision-maker takes of rejecting the null hypothesis when it is really true*. The guide provided by the statistical theory is that this probability must be ‘small’. Traditionally $\alpha = 0.05$ is selected for consumer research projects, $\alpha = 0.01$ for quality assurance and $\alpha = 0.10$ for political polling.

Step 3: Establish Critical or Rejection Region

The area under the sampling distribution curve of the test statistic is divided into two mutually exclusive regions (areas) as shown in Fig. 8.1. These regions are called the *acceptance region* and the *rejection* (or *critical*) *region*.

Figure 8.1
Areas of Acceptance and Rejection of H_0 (Two-Tailed Test)



The acceptance region is a *range of values* of the sample statistic spread around the *null hypothesized population parameter*. If values of the sample statistic fall within the limits of acceptance region, the null hypothesis is accepted, otherwise it is rejected.

The **rejection region** is the *range of sample statistic values* within which if values of the sample statistic falls (i.e. outside the limits of the acceptance region), then null hypothesis is rejected.

The value of the sample statistic that separates the regions of acceptance and rejection is called **critical value**.

The size of the rejection region is directly related to the level of precision to make decisions about a population parameter. Decision rules concerning null hypothesis are as follows:

- If $\text{prob}(H_0 \text{ is true}) \leq \alpha$, then reject H_0
- If $\text{prob}(H_0 \text{ is true}) > \alpha$, then accept H_0

In other words, if probability of H_0 being true is less than or equal to the significance level, α then reject H_0 , otherwise accept H_0 , i.e. the *level of significance α is used as the cut-off point which separates the area of acceptance from the area of rejection*.

Step 4: Select the Suitable Test of Significance or Test Statistic

The tests of significance or test statistic are classified into two categories: *parametric and nonparametric tests*. Parametric tests are more powerful because their data are derived from interval and ratio measurements. Nonparametric tests are used to test hypotheses with nominal and ordinal data. Parametric techniques are the tests of choice provided certain assumptions are met. Assumptions for parametric tests are as follows:

- (i) The selection of any element (or member) from the population should not affect the chance for any other to be included in the sample to be drawn from the population.
- (ii) The samples should be drawn from normally distributed populations.
- (iii) Populations under study should have equal variances.

Nonparametric tests have few assumptions and do not specify normally distributed populations or homogeneity of variance.

Selection of a test. For choosing a particular test of significance following three factors are considered:

- Whether the test involves one sample, two samples, or k samples?
- Whether two or more samples used are independent or related?
- Is the measurement scale nominal, ordinal, interval, or ratio?

Further, it is also important to know: (i) sample size, (ii) the number of samples, and their size, (iii) whether data have been weighted. Such questions help in selecting an appropriate test statistic.

One-sample tests are used for single sample and to test the hypothesis that it comes from a specified population. The following questions need to be answered before using one sample tests:

- Is there a difference between observed frequencies and the expected frequencies based on a statistical theory?
- Is there a difference between observed and expected proportions?
- Is it reasonable to conclude that a sample is drawn from a population with some specified distribution (normal, Poisson, and so on)?
- Is there a significant difference between some measures of central tendency and its population parameter?

The value of test statistic is calculated from the distribution of sample statistic by using the following formula

$$\text{Test statistic} = \frac{\text{Value of sample statistic} - \text{Value of hypothesized population parameter}}{\text{Standard error of the sample statistic}}$$

The choice of a probability distribution of a sample statistic is guided by the sample size n and the value of population standard deviation σ as shown in Table 8.1 and Fig 8.2.

Table 8.1: Choice of Probability Distribution

Sample Size n	Population Standard Deviation σ	
	Known	Unknown
• $n > 30$	Normal distribution	Normal distribution
• $n \leq 30$, population being assumed normal	Normal distribution	t -distribution

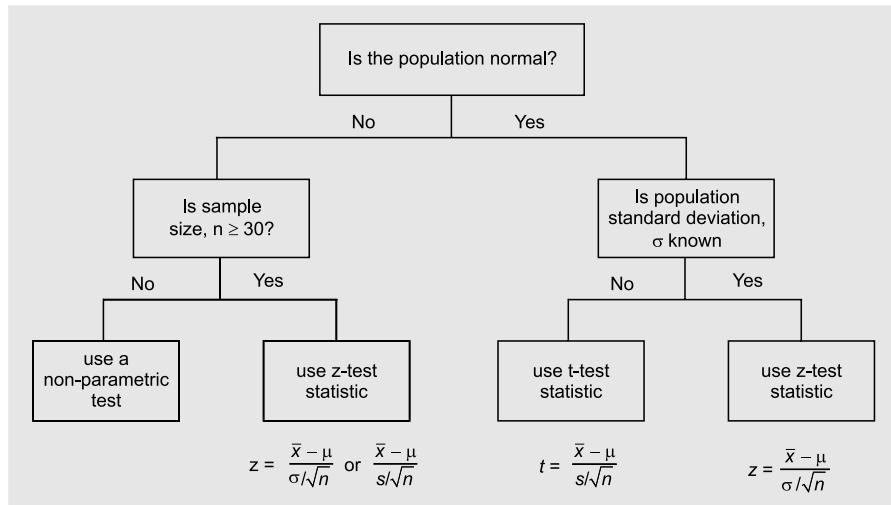


Figure 8.2
Choice of the Test Statistic

Step 5: Formulate a Decision Rule to Accept Null Hypothesis

Compare the calculated value of the test statistic with the critical value (also called *standard table value* of test statistic). The decision rules for null hypothesis are as follows:

- Accept H_0 if the test statistic value falls within the area of acceptance.
- Reject otherwise

In other words, if the calculated absolute value of a test statistic is less than or equal to its critical (or table) value, then accept the null hypothesis, otherwise reject it.

8.5 DIRECTION OF THE HYPOTHESIS TEST

The location of rejection region (or area) under the sampling distribution curve determines the direction of the hypothesis test, i.e. either lower tailed or upper tailed of the sampling distribution of relevant sample statistic being tested. It indicates the range of sample statistic values that would lead to a rejection of the null hypothesis. Figure 8.1, 8.3(a) and 8.3(b) illustrate the acceptance region and rejection region about a null hypothesized population mean, μ value for three different ways of formulating the null hypothesis..

- (i) Null hypothesis and alternative hypothesis stated as

$$H_0 : \mu = \mu_0 \text{ and } H_1 : \mu \neq \mu_0$$

imply that the sample statistic values which are either significantly smaller than or greater than the null hypothesized population mean, μ_0 value will lead to rejection of the null hypothesis. Hence, it is necessary to keep the rejection region at 'both tails' of the sampling distribution of the sample statistic. This type of test is called *two-tailed test* or *non-directional test* as shown in Fig. 8.1. If the significance level for the test is α per cent, then rejection region equal to $\alpha/2$ per cent is kept in each tail of the sampling distribution.

- (ii) Null hypothesis and alternative hypothesis stated as

$$H_0 : \mu \leq \mu_0 \text{ and } H_1 : \mu > \mu_0 \text{ (Right-tailed test)}$$

$$\text{or } H_0 : \mu \geq \mu_0 \text{ and } H_1 : \mu < \mu_0 \text{ (Left-tailed test)}$$

imply that the value of sample statistic is either 'higher than (or above)' or 'lower than (or below)' than the hypothesized population mean, μ_0 value. This lead to the rejection of null hypothesis for significant deviation from the specified value μ_0 in one direction (or tail) of the sampling distribution. Thus, the entire rejection region corresponding to the level of significance, α per cent, lies only in one tail of the sampling distribution of the sample statistic, as shown in Fig. 8.3(a) and (b). This type of test is called **one-tailed test** or *directional test*.

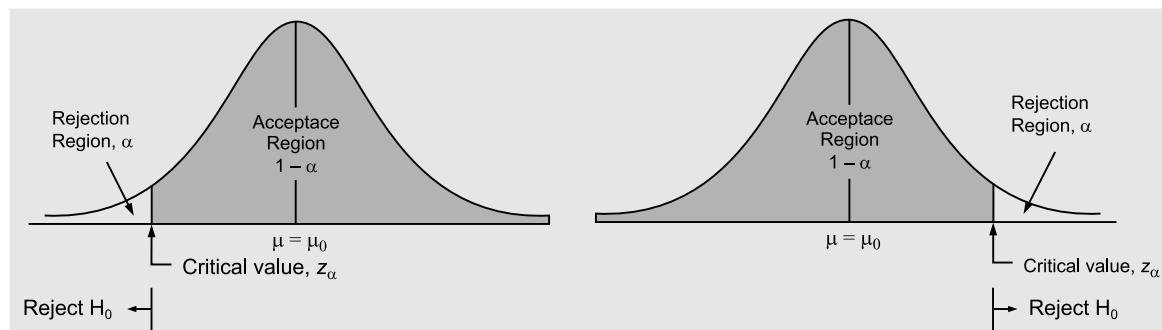


Fig. 8.3: (a) $H_0 : \mu \geq \mu_0; H_1 : \mu < \mu_0$, Left-tailed Test Fig. 8.3: (b) $H_0 : \mu \leq \mu_0; H_1 : \mu > \mu_0$, Right-tailed Test,

8.6 ERRORS IN HYPOTHESIS TESTING

Ideally the **hypothesis testing** procedure should lead to the acceptance of the null hypothesis H_0 when it is true, and the rejection of H_0 when it is not. However, the correct decision is not always possible. Since the decision to reject or accept a hypothesis is based on sample data, there is a possibility of an incorrect decision or error. A decision-maker may commit two types of errors while testing a null hypothesis. The two types of errors that can be made in any hypothesis testing are shown in Table 8.2.

Table 8.2: Errors in Hypothesis Testing

Decision	State of Nature	
	H_0 is True	H_0 is False
Accept H_0	Correct decision with confidence $(1 - \alpha)$	Type II error (β)
Reject H_0	Type I error (α)	Correct decision $(1 - \beta)$

Type I Error This is the *probability of rejecting the null hypothesis when it is true* and some alternative hypothesis is wrong. The probability of making a Type I error is denoted by the symbol α . It is represented by the area under the sampling distribution curve over the region of rejection.

The probability of making a Type I error, is referred to as the **level of significance**. The probability level of this error is decided by the decision-maker before the hypothesis test is performed and is based on his tolerance in terms of risk of rejecting the true null hypothesis. The risk of making Type I error depends on the cost and/or goodwill loss. The complement $(1 - \alpha)$ of the probability of Type I error measures the probability level of not rejecting a true null hypothesis. It is also referred to as *confidence level*.

Type II Error This is the *probability of accepting the null hypothesis when it is false* and some alternative hypothesis is true. The probability of making a Type II is denoted by the symbol β .

The probability of Type II error varies with the actual values of the population parameter being tested when null hypothesis H_0 is false. The probability of committing a Type II error depends on five factors: (i) the actual value of the population parameter, being tested, (ii) the level of significance selected, (iii) type of test (one or two tailed test) used to evaluate the null hypothesis, (iv) the sample standard deviation (also called standard error) and (v) the size of sample.

A summary of certain critical values at various significance levels for test statistic z is given in Table 8.3.

Table 8.3: Summary of Certain Critical Values for Sample Statistic z

Rejection Region	Level of Significance, α per cent			
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.005$
One-tailed region	± 1.285	± 1.645	± 2.33	± 2.58
Two-tailed region	± 1.645	± 1.96	± 2.58	± 2.81

8.6.1 Power of a Statistical Test

Another way of evaluating the goodness of a statistical test is to look at the complement of Type II error, which is stated as:

$$1 - \beta = P(\text{reject } H_0 \text{ when } H_1 \text{ is true})$$

The complement $1 - \beta$ of β , i.e. the probability of Type-II error, is called the *power of a statistical test* because it measures the probability of rejecting H_0 when it is true.

For example, suppose null and alternative hypotheses are stated as.

$$H_0: \mu = 80 \text{ and } H_1: \mu = 80$$

Hypothesis testing: The process of testing a statement or belief about a population parameter by the use of information collected from a sample(s).

Type I error: The probability of rejecting a true null hypothesis.

Level of significance: The probability of rejecting a true null hypothesis due to sampling error.

Type II error: The probability of accepting a false null hypothesis.

Power of a test: The ability (probability) of a test to reject the null hypothesis when it is false.

Often, when the null hypothesis is false, another alternative value of the population mean, μ is unknown. So for each of the possible values of the population mean μ , the probability of committing Type II error for several possible values of μ is required to be calculated.

Suppose a sample of size $n = 50$ is drawn from the given population to compute the probability of committing a Type II error for a specific alternative value of the population mean, μ . Let sample mean so obtained be $\bar{x} = 71$ with a standard deviation, $s = 21$. For significance level, $\alpha = 0.05$ and a two-tailed test, the table value of $z_{0.05} = \pm 1.96$. But the deserved value from sample data is

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{71 - 80}{21/\sqrt{50}} = -3.03$$

Since $z_{\text{cal}} = -3.03$ value falls in the rejection region, the null hypothesis H_0 is rejected. The rejection of null hypothesis, leads to either make a correct decision or commit a Type II error. If the population mean is actually 75 instead of 80, then the probability of committing a Type II error is determined by computing a critical region for the mean \bar{x}_c . This value is used as the cutoff point between the area of acceptance and the area of rejection. If for any sample mean so obtained is less than (or greater than for right-tail rejection region), \bar{x}_c , then the null hypothesis is rejected. Solving for the critical value of mean gives

$$\begin{aligned} z_c &= \frac{\bar{x}_c - \mu}{\sigma_{\bar{x}}} \text{ or } \pm 1.96 = \frac{\bar{x}_c - 80}{21/\sqrt{50}} \\ \bar{x}_c &= 80 \pm 5.82 \text{ or } 74.18 \text{ to } 85.82 \end{aligned}$$

If $\mu = 75$, then probability of accepting the false null hypothesis $H_0 : \mu = 80$ when critical value is falling in the range $\bar{x}_c = 74.18$ to 85.82 is calculated as follows:

$$z_1 = \frac{74.18 - 75}{21/\sqrt{50}} = -0.276$$

The area under normal curve for $z_1 = -0.276$ is 0.1064.

$$z_2 = \frac{85.82 - 75}{21/\sqrt{50}} = 3.643$$

The area under normal curve for $z_2 = 3.643$ is 0.4995

Thus the probability of committing a Type II error (β) falls in the region:

$$\beta = P(74.18 < \bar{x}_c < 85.82) = 0.1064 + 0.4995 = 0.6059$$

The total probability 0.6059 of committing a Type II error (β) is the area to the right of $\bar{x}_c = 74.18$ in the distribution. Hence the power of the test is $1 - \beta = 1 - 0.6059 = 0.3941$ as shown in Fig. 8.4(b).

Figure 8.4 (a)
Sampling distribution with
 $H_0 : \mu = 80$

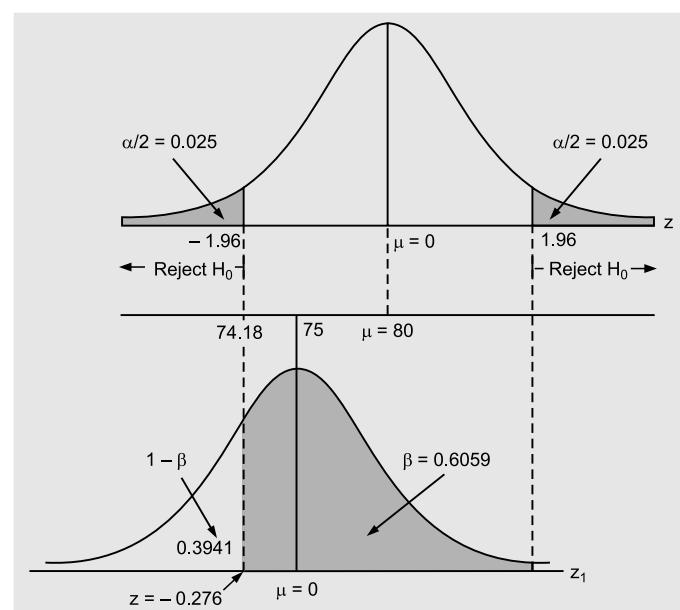


Figure 8.4 (b)
Sampling distribution with
 $H_0 : \mu = 75$

To keep α or β low depends on which type of error is more costly. However, if both types of errors are costly, then to keep both α and β low, then inferences can be made more reliable by reducing the variability of observations. It is preferred to have large sample size and a low α value.

Few relations between two errors α and β , the power of a test $1 - \beta$, and the sample size n are stated below:

- (i) If α (the sum of the two tail areas in the curve) is increased, the shaded area corresponding to β gets smaller, and vice versa.
- (ii) The β value can be increased for a fixed α , by increasing the sample size n .

Special Case: Suppose hypotheses are defined as:

$$H_0 : \mu = 80 \text{ and } H_1 : \mu < 80$$

Given $n = 50$, $s = 21$ and $\bar{x} = 71$. For $\alpha = 0.05$ and left-tailed test, the table value $z_{0.05} = -1.645$. The observed z value from sample data is

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{71 - 80}{21/\sqrt{50}} = -3.03$$

The critical value of the sample mean \bar{x}_c for a given population mean $\mu = 80$ is given by:

$$\begin{aligned} z_c &= \frac{\bar{x}_c - \mu}{\sigma_{\bar{x}}} \text{ or } -1.645 = \frac{\bar{x}_c - 80}{21/\sqrt{50}} \\ \bar{x}_c &= 75.115 \end{aligned}$$

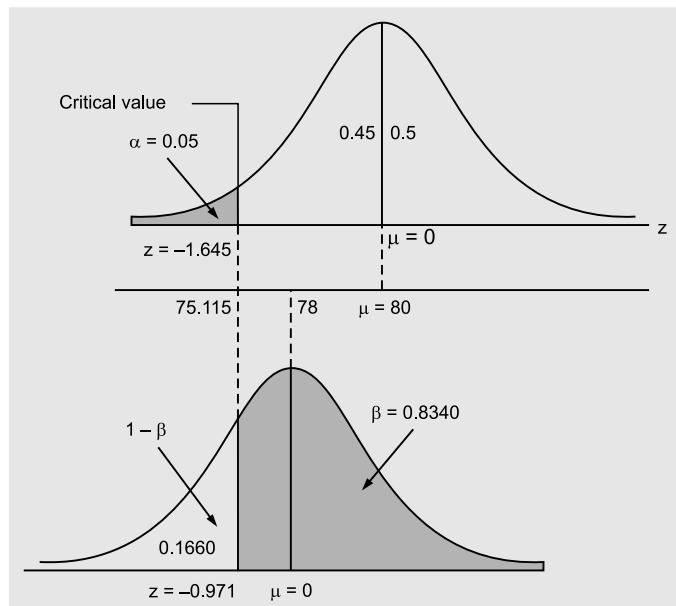


Figure 8.5 (a)
Sampling distribution with
 $H_0 : \mu = 80$

Figure 8.5 (b)
Sampling distribution with
 $H_0 : \mu = 78$

Fig. 8.5 (a) shows that the distribution of values that contains critical value of mean $\bar{x}_c = 75.115$ and below which H_0 will be rejected. Fig 8.5(b) shows the distribution of values when the alternative population mean value $\mu = 78$ is true. If H_0 is false, it is not possible to reject null hypothesis H_0 whenever sample mean is in the acceptance region, $\bar{x} \geq 75.115$. Thus critical value is computed by extending it and solved for the area to the right of \bar{x}_c as follows:

$$z_1 = \frac{\bar{x}_c - \mu}{\sigma_{\bar{x}}} = \frac{75.115 - 78}{21/\sqrt{50}} = -0.971$$

This value of z yields an area of 0.3340 under the normal curve. Thus the probability $= 0.3340 + 0.5000 = 0.8340$ of committing a Type II error is all the area to right of $\bar{x}_c = 75.115$.

Remark In general, if alternative value of population mean μ is relatively more than its hypothesized value, then probability of committing a Type II error is smaller compared to the case when the alternative value is close to the hypothesized value. The probability of committing a Type II error decreases as alternative values are greater than the hypothesized mean of the population.

Conceptual Questions 8A

1. Discuss the difference in purpose between the estimation of parameters and the testing of statistical hypothesis.
2. Describe the various steps involved in testing of hypothesis. What is the role of standard error in testing of hypothesis?
[Delhi Univ., M.Com, 1999]
3. What do you understand by null hypothesis and level of significance? Explain with the help of one example.
[HP Univ., MBA, 1996]
4. What is a test statistic? How is it used in hypothesis testing?
5. Define the term 'level of significance'. How is it related to the probability of committing a Type I error?
 - (a) Explain the general steps needed to carry out a test of any hypothesis.
 - (b) Explain clearly the procedure of testing hypothesis. Also point out the assumptions in hypothesis testing in large samples. [Kurukshetra Univ., MPWL, 1997]
6. This is always a trade-off between Type I and Type II errors. Discuss. [Delhi Univ., MCom, 1999]
7. When should a one-tailed alternative hypothesis be used? Under what circumstances is each type of test used?
8. Explain the general procedure for determining a critical value needed to perform a test of a hypothesis.
9. What is meant by the terms hypothesis and a test of a hypothesis?
10. Define the standard error of a statistic. How is it helpful in testing of hypothesis and decision-making?
11. Define the terms 'decision rule' and 'critical value'. What is the relationship between these terms?
12. (a) How is power related to the probability of making a Type II error?
 (b) What is the power of a hypothesis test? Why is it important.
 (c) How can the power of a hypothesis test be increased without increasing the sample size?
13. Write short notes on the following:
 - (a) Acceptance and rejection regions
 - (b) Type I and Type II errors
 - (c) Null and alternative hypotheses
 - (d) One-tailed and two-tailed tests
14. When planning a hypothesis test, what should be done if probabilities of both Type I and Type II are to be small

8.7 HYPOTHESIS TESTING FOR POPULATION PARAMETERS WITH LARGE SAMPLES

Hypothesis testing involving large samples ($n > 30$) is based on the assumption that the population from which the sample is drawn has a normal distribution. Consequently the sampling distribution of mean \bar{x} is also normal. Even if the population does not have a normal distribution, the sampling distribution of mean \bar{x} is assumed to be normal due to the central limit theorem because the sample size is large.

8.7.1 Hypothesis Testing for Single Population Mean

Two-tailed Test Let μ_0 be the hypothesized value of the population mean to be tested. For this the null and alternative hypotheses for two-tailed test are defined as:

$$H_0 : \mu = \mu_0 \quad \text{or} \quad \mu - \mu_0 = 0$$

and

$$H_1 : \mu \neq \mu_0$$

If standard deviation σ of the population is known, then based on the central limit theorem, the sampling distribution of mean \bar{x} would follow the standard normal distribution for a large sample size. The z-test statistic is given by

$$\text{Test-statistic: } z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

In this formula, the numerator $\bar{x} - \mu$, measures how far (in an absolute sense) the observed sample mean \bar{x} is from the hypothesized mean μ . The denominator $\sigma_{\bar{x}}$ is the standard error of the mean, so the z-test statistic represents how many standard errors \bar{x} is from μ .

If the population standard deviation σ is not known, then a sample standard deviation s is used to estimate σ . The value of the z-test statistic is given by

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

The two rejection areas in two-tailed test are determined so that half the level of significance, $\alpha/2$ appears in each tail of the distribution of mean. Hence $z_{\alpha/2}$ represents

the standardized normal variate corresponding to $\alpha/2$ in both the tails of normal curve as shown in Fig 8.1. The decision rule based on sample mean for the two-tailed test takes the form:

- Reject H_0 if $z_{\text{cal}} \leq -z_{\alpha/2}$ or $z_{\text{cal}} \geq z_{\alpha/2}$
- Accept H_0 if $-z_{\alpha/2} < z < z_{\alpha/2}$

where $z_{\alpha/2}$ is the table value (also called CV, critical value) of z at a chosen level of significance α .

Left-tailed Test Large sample ($n > 30$) hypothesis testing about a population mean for a left-tailed test is of the form

$$H_0 : \mu \geq \mu_0 \quad \text{and} \quad H_1 : \mu < \mu_0$$

$$\text{Test statistic: } z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- Decision rule:
- Reject H_0 if $z_{\text{cal}} \leq -z_{\alpha}$ (Table value of z at α)
 - Accept H_0 if $z_{\alpha/2} > z_{\text{cal}}$

Right-tailed Test Large sample ($n > 30$) hypothesis testing about a population mean for a right-tailed test is of the form

$$H_0 : \mu \leq \mu_0 \text{ and } H_1 : \mu > \mu_0 \text{ (Right-tailed test)}$$

$$\text{Test statistic: } z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- Decision rule:
- Reject H_0 if $z_{\text{cal}} \geq z_{\alpha}$ (Table value of z at α)
 - Accept H_0 if $z_{\text{cal}} < z_{\alpha}$

8.7.2 Relationship between Interval Estimation and Hypothesis Testing

Consider following statements of null and alternative hypothesis:

- $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$ (Two-tailed test)
- $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$ (Right-tailed test)
- $H_0 : \mu \geq \mu_0$ and $H_1 : \mu < \mu_0$ (Left-tailed test)

The following are the confidence intervals in all above three cases where hypothesized value μ_0 of population mean, μ is likely to fall. Accordingly, the decision to accept or reject the null hypothesis will be taken.

Two-tailed test Two critical values CV_1 and CV_2 one for each tail of the sampling distribution is computed as follows:

(a) Known σ

Normal population : Any sample size, n

Any population : Large sample size n

$$CV_1 = \mu_0 - z_{\alpha/2} \sigma_{\bar{x}}$$

$$CV_2 = \mu_0 + z_{\alpha/2} \sigma_{\bar{x}}$$

$$\text{where } \sigma_{\bar{x}} = \sigma/\sqrt{n}$$

(b) Unknown σ

Any population : Large sample size, n

$$CV_1 = \mu_0 - z_{\alpha/2} s_{\bar{x}}$$

$$CV_2 = \mu_0 + z_{\alpha/2} s_{\bar{x}}$$

$$s_{\bar{x}} = s/\sqrt{n}$$

Two-tailed test: The test of a null hypothesis which can be rejected when the sample statistic is in either extreme end of the sampling distribution.

- Decision rule:**
- Reject H_0 when $\bar{x} \leq CV_1$ or $\bar{x} \geq CV_2$
 - Accept H_0 when $CV_1 < \bar{x} < CV_2$

Left-tailed test The critical value for left tail of the sampling distribution is computed as follows:

(a) Known σ	(b) Unknown σ
Normal population : Any sample size, n	Any population : Large sample size, n
Any population : Large sample size, n	
$CV = \mu_0 - z_\alpha \sigma_{\bar{x}}$	$CV = \mu_0 - z_\alpha s_{\bar{x}}$

- Decision rule:**
- Reject H_0 when $\bar{x} \leq CV$
 - Accept H_0 when $\bar{x} > CV$

Right-tailed test The critical value for right tail of the sampling distribution is computed as follows:

(a) Known σ	(b) Unknown σ
Normal population : Any sample size, n	Any population : Large sample size, n
Any population : Large sample size, n	
$CV = \mu_0 + z_\alpha \sigma_{\bar{x}}$	$CV = \mu_0 + z_\alpha s_{\bar{x}}$

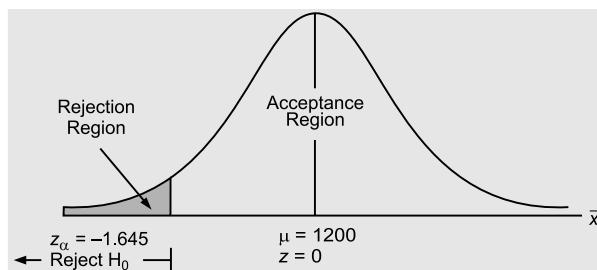
- Decision rule :
- Reject H_0 when $\bar{x} \geq CV$
 - Accept H_0 when $\bar{x} < CV$

Example 8.1: Individual filing of income tax returns prior to 30 June had an average refund of Rs 1200. Consider the population of 'last minute' filers who file their returns during the last week of June. For a random sample of 400 individuals who filed a return between 25 and 30 June, the sample mean refund was Rs 1054 and the sample standard deviation was Rs 1600. Using 5 per cent level of significance, test the belief that the individuals who wait until the last week of June to file their returns to get a higher refund than early filers.

Solution: Since population standard deviation is not given, the standard error must be estimated with $s_{\bar{x}}$. Let us take the null hypothesis H_0 that the individuals who wait until the last week of June to file their returns get a higher return than the early filers, that is,

$$H_0 : \mu \geq 1200 \quad \text{and} \quad H_1 : \mu < 1200 \quad (\text{Left-tailed test})$$

Figure 8.6



Given, $n = 400$, $s = 1600$, $\bar{x} = 1054$, $\alpha = 5\%$. Thus using the z -test statistic

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1054 - 1200}{1600/\sqrt{400}} = -\frac{146}{80} = -1.825$$

Since the calculated value $z_{\text{cal}} = -1.825$ is less than its critical value $z_\alpha = -1.645$ at $\alpha = 0.05$ level of significance, the null hypothesis, H_0 is rejected, as shown in Fig. 8.6. Hence, we conclude that individuals who wait until the last week of June are likely to receive a refund of less than Rs 1200.

$$\begin{aligned} \text{Alternative approach: } CV &= \mu_0 - z_\alpha \sigma_{\bar{x}} = 1200 - 1.645 \times (1600/\sqrt{400}) \\ &= 1200 - 131.6 = 1068.4 \end{aligned}$$

Since $\bar{x} (= 1054) < CV (= 1068.4)$, the null hypothesis H_0 is rejected

Example 8.2: A packaging device is set to fill detergent powder packets with a mean weight of 5 kg, with a standard deviation of 0.21 kg. The weight of packets can be assumed to be normally distributed. The weight of packets is known to drift upwards over a period of time due to machine fault, which is not tolerable. A random sample of 100 packets is taken and weighed. This sample has a mean weight of 5.03 kg. Can we conclude that the mean weight produced by the machine has increased? Use a 5 per cent level of significance.

Solution: Let us take the null hypothesis H_0 that mean weight has increased, that is,

$$H_0 : \mu \geq 5 \text{ and } H_1 : \mu < 5$$

Given $n = 100$, $\bar{x} = 5.03$ kg, $\sigma = 0.21$ kg and $\alpha = 5$ per cent. Thus using the z-test statistic

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{5.03 - 5}{0.21/\sqrt{100}} = \frac{0.03}{0.021} = 1.428$$

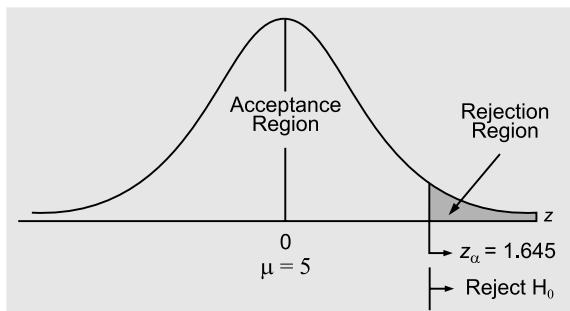


Figure 8.7

Since calculated value $z_{\text{cal}} = 1.428$ is less than its critical value $z_\alpha = 1.645$ at $\alpha = 0.05$, the null hypothesis, H_0 is accepted as shown in Fig. 8.4. Hence we conclude that mean weight is likely to be more than 5 kg.

$$\begin{aligned} \text{Alternative approach: } CV &= \mu_0 + z_\alpha \sigma_{\bar{x}} = 5 + 1.645 \times (0.21/\sqrt{100}) \\ &= 5 + 0.034 = 5.034 \end{aligned}$$

Since $\bar{x} (= 5.03) < CV (= 5.034)$, H_0 is accepted.

Example 8.3: The mean life time of a sample of 400 fluorescent light bulbs produced by a company is found to be 1600 hours with a standard deviation of 150 hours. Test the hypothesis that the mean life time of the bulbs produced in general is higher than the mean life of 1570 hours at $\alpha = 0.01$ level of significance.

Solution: Let us take the null hypothesis that mean life time of bulbs is not more than 1570 hours, that is

$$H_0 : \mu \leq 1570 \text{ and } H_1 : \mu > 1570 \text{ (Right-tailed test)}$$

Given $n = 400$, $\bar{x} = 1600$ hours, $s = 150$ hrs and $\alpha = 0.01$. Thus using the z-test statistic.

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1600 - 1570}{150/\sqrt{400}} = \frac{30}{7.5} = 4$$

Since the calculated value $z_{\text{cal}} = 4$ is more than its critical value $z_\alpha = \pm 2.33$, the H_0 is rejected. Hence, we conclude that the mean lifetime of bulbs produced by the company may be higher than 1570 hours.

$$\begin{aligned} \text{Alternatively approach: } CV &= \mu_0 + z_\alpha s_{\bar{x}} = 1570 + 2.33 \times (150/\sqrt{400}) \\ &= 1570 + 17.475 = 1587.475 \end{aligned}$$

Since $\bar{x} (= 1600) > CV (= 1587.47)$, the null hypothesis H_0 is rejected.

Example 8.4: A continuous manufacturing process of steel rods is said to be in 'state of control' and produces acceptable rods if the mean diameter of all rods produced is 2 inches. Although the process standard deviation exhibits stability over time with standard deviation, $\sigma = 0.01$ inch. The process mean may vary due to operator error or problems of process adjustment. Periodically, random samples of 100 rods are selected to determine whether the process is producing acceptable rods. If the result of a test indicates that the process is out of control, it is stopped and the source of trouble is sought. Otherwise, it

is allowed to continue operating. A random sample of 100 rods is selected resulting in a mean of 2.1 inches. Test the hypothesis to determine whether the process be continued.

Solution: Since rods that are either too narrow or too wide are unacceptable, the low values and high values of the sample mean lead to the rejection of the null hypothesis. Consider the null hypothesis H_0 , that the process may be allowed to continue when diameter is 2 inches. Consequently, rejection region is on both tails of the sampling distribution. The null and alternative hypotheses are stated as follows:

$$H_0 : \mu = 2 \text{ inches, (continue process)}$$

$$H_1 : \mu \neq 2 \text{ inches, (stop the process)}$$

Given $n = 100$, $\bar{x} = 2.1$, $\sigma = 0.01$, $\alpha = 0.01$. Using the z-test statistic

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{2.1 - 2}{0.01/\sqrt{100}} = \frac{0.1}{0.001} = 100$$

Since $z_{\text{cal}} = 100$ value is more than its critical value $z_{\alpha/2} = 2.58$ at $\alpha = 0.01$, the null hypothesis, H_0 is rejected. Thus stop the process in order to determine the source of trouble.

$$\begin{aligned} \text{Alternative approach: } CV_1 &= \mu_0 - z_{\alpha/2} \sigma_{\bar{x}} = \mu_0 - z_{\alpha/2} (\sigma/\sqrt{n}) \\ &= 2 - 2.58 \times (0.01/\sqrt{100}) = 2 - 0.003 = 1.997 \\ CV_2 &= \mu_0 + z_{\alpha/2} \sigma_{\bar{x}} = 2 + 2.58 \times (0.01/\sqrt{100}) \\ &= 2 + 0.003 = 2.003 \end{aligned}$$

Since $\bar{x} (= 2.1) \geq CV_2 (= 2.003)$, the null hypothesis is rejected.

Example 8.5: An ambulance service claims that it takes, on the average 8.9 minutes to reach its destination in emergency calls. To check on this claim, the agency which licenses ambulance services has then timed on 50 emergency calls, getting a mean of 9.3 minutes with a standard deviation of 1.8 minutes. Does this constitute evidence that the figure claimed is too low at the 1 per cent significance level?

Solution: Let us consider the null hypothesis H_0 that 'the claim is same as observed' and alternative hypothesis is 'claim is different than observed'. These two hypotheses are written as:

$$H_0 : \mu = 8.9 \text{ and } H_1 : \mu \neq 8.9$$

Given $n = 50$, $\bar{x} = 9.3$, and $s = 1.8$. Using the z-test statistic, we get

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{9.3 - 8.9}{1.8/\sqrt{50}} = \frac{0.4}{0.254} = 1.574$$

Since $z_{\text{cal}} = 1.574$ is less than its critical value $z_{\alpha/2} = \pm 2.58$, at $\alpha = 0.01$, the null hypothesis is accepted. Thus, there is no difference between the average time observed and claimed.

8.7.3 p-Value Approach to Test Hypothesis of Single Population Mean

p-value: The probability of getting the sample statistic or a more extreme value, when null hypothesis is true.

The **p-value** is another approach for hypothesis testing of population mean based on a large sample. This is often referred to as the *observed significance level*, that is, the smallest significance level α for which null hypothesis H_0 can be rejected. It is the actual risk of committing Type I error when the null hypothesis, H_0 is rejected based on the observed value of the test statistic. The *p-value* measures the strength of evidence against H_0 , i.e. a *p-value* is a way to express the likelihood that H_0 is not true. In other words, *p-value* is the probability of observing a sample value as extreme as or more extreme than, the value of test statistic, given that the null hypothesis H_0 is true. The advantage of this approach is that the *p-value* can be compared directly to the level of significance α .

The decision rules for accepting or rejecting a null hypothesis based on the *p-value* are as follows:

- (i) For a left-tailed test, the *p-value* is the area to the left of the calculated value of the test statistic. For instance, if $z_{\text{cal}} = -1.76$, then the area to the left of it is $0.5000 - 0.4608 = 0.0392$ or the *p-value* is 3.92 per cent

- (ii) For right-tailed test, the p -value is the area to the right of the calculated value of the test statistic. For instance, if $z_{\text{cal}} = +2.00$, then the area to the right of it is $0.5000 - 0.4772 = 0.0228$, or the p value is 2.28 per cent.

Thus the decision rules for left-tailed test and right-tailed test are as under.

- Reject H_0 if p -value $\leq \alpha$
- Accept H_0 if p -value $> \alpha$

- (iii) For a two-tailed test, the p -value is twice the tail area. If the calculated value of the test statistic falls in the left tail (or right tail), then the area to the left (or right) of the calculated value is multiplied by 2.

Example 8.6: An auto company decided to introduce a new six cylinder car whose mean petrol consumption is claimed to be lower than that of the existing auto engine. It was found that the mean petrol consumption for 50 cars was 10 km per litre with a standard deviation of 3.5 km per litre. Test for the company at 5 per cent level of significance, the claim that in the new car petrol consumption is 9.5 km per litre on the average. [HP Univ., MBA, 1989]

Solution: Let us assume the null hypothesis H_0 that there is no significant difference between the company's claim and the sample average value, that is,

$$H_0 : \mu = 9.5 \text{ km/litre} \text{ and } H_1 : \mu \neq 9.5 \text{ km/litre}$$

Given $\bar{x} = 10$, $n = 50$, $s = 3.5$, and $z_{\alpha/2} = 1.96$ at $\alpha = 0.05$ level of significance. Thus using the z -test statistic

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{10 - 9.5}{3.5/\sqrt{50}} = 1.010$$

Since $z_{\text{cal}} = 1.010$ is less than its critical value $z_{\alpha/2} = 1.96$ at $\alpha = 0.05$ level of significance, the null hypothesis is accepted. Hence we can conclude that the new car's petrol consumption is 9.5 km/litre.

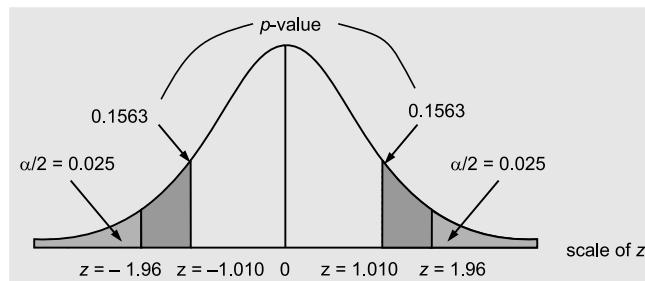


Figure 8.8

The p -value approach The null hypothesis accepted H_0 because $z_{\text{cal}} = 1.010$ lies in the acceptance region. The probability of finding $z_{\text{cal}} = 1.010$ or more is 0.3437 (from normal table). The p -value is the area to the right as well as left of the calculated value of z -test statistic (for two-tailed test). Since $z_{\text{cal}} = 1.010$, then the area to its right is $0.5000 - 0.3437 = 0.1563$ as shown in Fig. 8.8.

Since it is the two-tailed test, p -value becomes $2(0.1563) = 0.3126$. Since $0.3126 > \alpha = 0.05$, null hypothesis H_0 is accepted.

8.7.4 Hypothesis Testing for Difference between Two Population Means

If we have two independent populations each having its mean and standard deviation as:

Population	Mean	Standard Deviation
1	μ_1	σ_1
2	μ_2	σ_2

then we can extend the hypothesis testing concepts developed in the previous section to test whether there is any significant difference between the means of these populations.

Let two independent random samples of large size n_1 and n_2 be drawn from the first and second population, respectively. Let the sample means so calculated be \bar{x}_1 and \bar{x}_2 .

The z-test statistic used to determine the difference between the population means ($\mu_1 - \mu_2$) is based on the difference between the sample means ($\bar{x}_1 - \bar{x}_2$) because sampling distribution of $\bar{x}_1 - \bar{x}_2$ has the property $E(\bar{x}_1 - \bar{x}_2) = (\mu_1 - \mu_2)$. This test statistic will follow the standard normal distribution for a large sample due to the central limit theorem. The z-test statistic is

$$\text{Test statistic: } z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where $\sigma_{\bar{x}_1 - \bar{x}_2}$ = standard error of the statistic ($\bar{x}_1 - \bar{x}_2$)

$\bar{x}_1 - \bar{x}_2$ = difference between two sample means, that is, sample statistic

$\mu_1 - \mu_2$ = difference between population means, that is, hypothesized population parameter

If $\sigma_1^2 = \sigma_2^2$, the above formula algebraically reduces to:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

If the standard deviations σ_1 and σ_2 of each of the populations are *not known*, then we may estimate the standard error of sampling distribution of the sample statistic $\bar{x}_1 - \bar{x}_2$ by substituting the sample standard deviations s_1 and s_2 as estimates of the population standard deviations. Under this condition, the standard error of $\bar{x}_1 - \bar{x}_2$ is estimated as:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The standard error of the *difference between standard deviation of sampling distribution* is given by

$$\sigma_{\sigma_1 - \sigma_2} = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$$

The null and alternative hypothesis are stated as:

$$\begin{array}{ll} \text{Null hypothesis} & : H_0 : \mu_1 - \mu_2 = d_0 \\ \text{Alternative hypothesis} & : \end{array}$$

One-tailed Test	Two-tailed Test
$H_1 : (\mu_1 - \mu_2) > d_0$	$H_1 : (\mu_1 - \mu_2) \neq d_0$
$H_1 : (\mu_1 - \mu_2) < d_0$	

where d_0 is some specified difference that is desired to be tested. If there is no difference between μ_1 and μ_2 , i.e. $\mu_1 = \mu_2$, then $d_0 = 0$.

Decision rule: Reject H_0 at a specified level of significance α when

One-tailed test	Two-tailed test
<ul style="list-style-type: none"> • $z_{\text{cal}} > z_\alpha$ [or $z < -z_\alpha$ when $H_1 : \mu_1 - \mu_2 < d_0$] • When $p\text{-value} < \alpha$ 	<ul style="list-style-type: none"> • $z_{\text{cal}} > z_{\alpha/2}$ or $z_{\text{cal}} < -z_{\alpha/2}$

Example 8.7: A firm believes that the tyres produced by process A on an average last longer than tyres produced by process B. To test this belief, random samples of tyres produced by the two processes were tested and the results are:

Process	Sample Size	Average Lifetime (in km)	Standard Deviation (in km)
A	50	22,400	1000
B	50	21,800	1000

Is there evidence at a 5 per cent level of significance that the firm is correct in its belief?

Solution: Let us take the null hypothesis that there is no significant difference in the average life of tyres produced by processes A and B, that is,

$$H_0 : \mu_1 = \mu_2 \text{ or } \mu_1 - \mu_2 = 0 \quad \text{and} \quad H_1 : \mu_1 \neq \mu_2$$

Given, $\bar{x}_1 = 22,400$ km, $\bar{x}_2 = 21,800$ km, $\sigma_1 = \sigma_2 = 1000$ km, and $n_1 = n_2 = 50$. Thus using the z-test statistic

$$\begin{aligned} z &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \\ &= \frac{22,400 - 21,800}{\sqrt{\frac{(1000)^2}{50} + \frac{(1000)^2}{50}}} = \frac{600}{\sqrt{20,000 + 20,000}} = \frac{600}{200} = 3 \end{aligned}$$

Since the calculated value $z_{\text{cal}} = 3$ is more than its critical value $z_{\alpha/2} = \pm 1.645$ at $\alpha = 0.05$ level of significance, therefore H_0 is rejected. Hence we can conclude that the tyres produced by process A last longer than those produced by process B.

The p-value approach:

$$\begin{aligned} p\text{-value} &= P(z > 3.00) + P(z < -3.00) = 2 P(z > 3.00) \\ &= 2(0.5000 - 0.4987) = 0.0026 \end{aligned}$$

Since p -value of 0.026 is less than specified significance level $\alpha = 0.05$, H_0 is rejected.

Example 8.8: An experiment was conducted to compare the mean time in days required to recover from a common cold for person given daily dose of 4 mg of vitamin C versus those who were not given a vitamin supplement. Suppose that 35 adults were randomly selected for each treatment category and that the mean recovery times and standard deviations for the two groups were as follows:

	Vitamin C	No Vitamin Supplement
Sample size	35	35
Sample mean	5.8	6.9
Sample standard deviation	1.2	2.9

Test the hypothesis that the use of vitamin C reduces the mean time required to recover from a common cold and its complications, at the level of significance $\alpha = 0.05$.

Solution: Let us take the null hypothesis that the use of vitamin C reduces the mean time required to recover from the common cold, that is

$$H_0 : (\mu_1 - \mu_2) \leq 0 \text{ and } H_1 : (\mu_1 - \mu_2) > 0$$

Given $n_1 = 35$, $\bar{x}_1 = 5.8$, $s_1 = 1.2$ and $n_2 = 35$, $\bar{x}_2 = 6.9$, $s_2 = 2.9$. The level of significance, $\alpha = 0.05$. Substituting these values into the formula for z-test statistic, we get

$$\begin{aligned} z &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{5.8 - 6.9}{\sqrt{\frac{(1.2)^2}{35} + \frac{(2.9)^2}{35}}} = \frac{-1.1}{\sqrt{0.041 + 0.240}} = -\frac{1.1}{0.530} = -2.065 \end{aligned}$$

Using a one-tailed test with significance level $\alpha = 0.05$, the critical value is $z_\alpha = 1.645$. Since $z_{\text{cal}} < z_\alpha (= 1.645)$, the null hypothesis H_0 is rejected. Hence we can conclude that the use of vitamin C does not reduce the mean time required to recover from the common cold.

Example 8.9: The Educational Testing Service conducted a study to investigate difference between the scores of female and male students on the Mathematics Aptitude Test. The

study identified a random sample of 562 female and 852 male students who had achieved the same high score on the mathematics portion of the test. That is, the female and male students viewed as having similar high ability in mathematics. The verbal scores for the two samples are given below:

	Female	Male
Sample mean	547	525
Sample standard deviation	83	78

Do the data support the conclusion that given populations of female and male students with similar high ability in mathematics, the female students will have a significantly high verbal ability? Test at $\alpha = 0.05$ significance level. What is your conclusion?

[Delhi Univ., MBA, 2003]

Solution: Let us take the null hypothesis that the female students have high level verbal ability, that is,

$$H_0 : (\mu_1 - \mu_2) \geq 0 \text{ and } H_1 : (\mu_1 - \mu_2) < 0$$

Given, for female students: $n_1 = 562$, $\bar{x}_1 = 547$, $s_1 = 83$, for male students: $n_2 = 852$, $\bar{x}_2 = 525$, $s_2 = 78$, and $\alpha = 0.05$.

Substituting these values into the z-test statistic, we get

$$\begin{aligned} z &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{547 - 525}{\sqrt{\frac{(83)^2}{562} + \frac{(78)^2}{852}}} \\ &= \frac{22}{\sqrt{12.258 + 7.140}} = \frac{22}{\sqrt{19.398}} = \frac{22}{4.404} = 4.995 \end{aligned}$$

Using a one-tailed test with $\alpha=0.05$ significance level, the critical value of z-test statistic is $z_\alpha = \pm 1.645$. Since $z_{\text{cal}} = 4.995$ is more than the critical value $z_\alpha = 1.645$, null hypothesis, H_0 is rejected. Hence, we conclude that there is no sufficient evidence to declare that difference between verbal ability of female and male students is significant.

Example 8.10: In a sample of 1000, the mean is 17.5 and the standard deviation is 2.5. In another sample of 800, the mean is 18 and the standard deviation is 2.7. Assuming that the samples are independent, discuss whether the two samples could have come from a population which have the same standard deviation.

[Saurashtra Univ., BCom, 1997]

Solution: Let us take the hypothesis that there is no significant difference in the standard deviations of the two samples, that is, $H_0 : \sigma_1 = \sigma_2$ and $H_1 : \sigma_1 \neq \sigma_2$.

Given, $\sigma_1 = 2.5$, $n_1 = 1000$ and $\sigma_2 = 2.7$, $n_2 = 800$. Thus we have

$$\begin{aligned} \text{Standard error, } \sigma_{\sigma_1 - \sigma_2} &= \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}} \\ &= \sqrt{\frac{(2.5)^2}{2000} + \frac{(2.7)^2}{1600}} = \sqrt{\frac{6.25}{2000} + \frac{7.29}{1600}} = 0.0876 \end{aligned}$$

Applying the z-test statistic, we have

$$z = \frac{\sigma_1 - \sigma_2}{\sigma_{\sigma_1 - \sigma_2}} = \frac{2.7 - 2.5}{0.0876} = \frac{0.2}{0.0876} = 2.283$$

Since the $z_{\text{cal}} = 2.283$ is more than its critical value $z = 1.96$ at $\alpha = 5$ per cent, the null hypothesis H_0 is rejected. Hence we conclude that the two samples have not come from a population which has the same standard deviation.

Example 8.11: The mean production of wheat from a sample of 100 fields is 200 lbs per acre with a standard deviation of 10 lbs. Another sample of 150 fields gives the mean at 220 lbs per acre with a standard deviation of 12 lbs. Assuming the standard deviation of the universe as 11 lbs, find at 1 per cent level of significance, whether the two results are consistent.

[Punjab Univ., MCom, Mangalore MBA, 1996]

Solution: Let us take the hypothesis that the two results are consistent, that is

$$H_0 : \sigma_1 = \sigma_2 \text{ and } H_1 : \sigma_1 \neq \sigma_2.$$

Given $\sigma_1 = \sigma_2 = 11$, $n_1 = 100$, $n_2 = 150$. Thus

$$\sigma_{\sigma_1 - \sigma_2} = \sqrt{\frac{\sigma^2}{2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{\frac{(11)^2}{2} \left(\frac{1}{100} + \frac{1}{150} \right)} = 1.004$$

Applying the z-test statistic we have

$$z = \frac{\sigma_1 - \sigma_2}{\sigma_{\sigma_1 - \sigma_2}} = \frac{10 - 12}{1.004} = -\frac{2}{1.004} = -1.992$$

Since the $z_{\text{cal}} = -1.992$ is more than its critical value $z = -2.58$ at $\alpha = 0.01$, the null hypothesis is accepted. Hence we conclude that the two results are likely to be consistent.

Self-Practice Problems 8A

- 8.1** The mean breaking strength of the cables supplied by a manufacturer is 1800 with a standard deviation of 100. By a new technique in the manufacturing process it is claimed that the breaking strength of the cables has increased. In order 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 at a 0.01 level of significance?
- 8.2** A sample of 100 households in a village was taken and the average income was found to be Rs 628 per month with a standard deviation of Rs 60 per month. Find the standard error of mean and determine 99 per cent confidence limits within which the income of all the people in this village are expected to lie. Also test the claim that the average income was Rs 640 per month.
- 8.3** A random sample of boots worn by 40 combat soldiers in a desert region showed an average life of 1.08 years with a standard deviation of 0.05. Under the standard conditions, the boots are known to have an average life of 1.28 years. Is there reason to assert at a level of significance of 0.05 that use in the desert causes the mean life of such boots to decrease?
- 8.4** An ambulance service claims that it takes, on an average, 8.9 minutes to reach its destination in emergency calls. To check on this claim, the agency which licenses ambulance services had them timed on 50 emergency calls, getting a mean of 9.3 minutes with a standard deviation of 1.8 minutes. At the level of significance of 0.05, does this constitute evidence that the figure claimed is too low?
- 8.5** A sample of 100 tyres is taken from a lot. The mean life of the tyres is found to be 39,350 km with a standard deviation of 3260 km. Could the sample come from a population with mean life of 40,000 km? Establish 99 per cent confidence limits within which the mean life of the tyres is expected to lie.

[Delhi Univ., BA(H) Eco., 1996]

- 8.6** A simple sample of the heights of 6400 Englishmen has a mean of 67.85 inches and a standard deviation of 2.56 inches, while a simple sample of heights of 1600

Austrians has a mean of 68.55 inches and a standard deviation of 2.52 inches. Do the data indicate that the Austrians are on the average taller than the Englishmen? Give reasons for your answer.

- [MD Univ., MCom, 1998; Kumaon Univ., MBA, 1999]
- 8.7** A man buys 50 electric bulbs of 'Philips' and 50 electric bulbs of 'HMT'. He finds that 'Philips' bulbs gave an average life of 1500 hours with a standard deviation of 60 hours and 'HMT' bulbs gave an average life of 1512 hours with a standard deviation of 80 hours. Is there a significant difference in the mean life of the two makes of bulbs?
- [MD Univ., MCom, 1998; Kumaon Univ., MBA, 1999]
- 8.8** Consider the following hypothesis:
- $$H_0 : \mu = 15 \text{ and } H_1 : \mu \neq 15$$
- A sample of 50 provided a sample mean of 14.2 and standard deviation of 5. Compute the p -value, and conclude about H_0 at the level of significance 0.02.
- 8.9** A product is manufactured in two ways. A pilot test on 64 items from each method indicates that the products of method 1 have a sample mean tensile strength of 106 lbs and a standard deviation of 12 lbs, whereas in method 2 the corresponding values of mean and standard deviation are 100 lbs and 10 lbs, respectively. Greater tensile strength in the product is preferable. Use an appropriate large sample test of 5 per cent level of significance to test whether or not method 1 is better for processing the product. State clearly the null hypothesis.
- [Delhi Univ., MBA, 2003]
- 8.10** Two types of new cars produced in India are tested for petrol mileage. One group consisting of 36 cars averaged 14 kms per litre. While the other group consisting of 72 cars averaged 12.5 kms per litre.
- What test-statistic is appropriate if $\sigma_1^2 = 1.5$ and $\sigma_2^2 = 2.0$?
 - Test whether there exists a significant difference in the petrol consumption of these two types of cars. (use $\alpha = 0.01$)
- [Roorkee Univ., MBA, 2000]

Hints and Answers

- 8.1** Let $H_0 : \mu = 1800$ and $H_1 : \mu \neq 1800$ (Two-tailed test)

Given $\bar{x} = 1850$, $n = 50$, $\sigma = 100$, $z_\alpha = \pm 2.58$ at $\alpha = 0.01$ level of significance

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{1850 - 1800}{100/\sqrt{50}} = 3.54$$

Since $z_{\text{cal}} (= 3.54) > z_\alpha (= 2.58)$, reject H_0 . The breaking strength of the cables of 1800 does not support the claim.

- 8.2** Given $n = 100$, $\sigma = 50$; $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{50}{\sqrt{100}} = 5$.

Confidence interval at 99% is: $\bar{x} \pm z_\alpha \sigma_{\bar{x}} = 628 \pm 2.58 (5) = 628 \pm 12.9$; $615.1 \leq \mu \leq 640.9$

Since hypothesized population mean $\mu = 640$ lies in the this interval, H_0 is accepted.

- 8.3** Let $H_0 : \mu = 1.28$ and $H_1 : \mu < 1.28$ (One-tailed test)

Given $n = 40$, $\bar{x} = 1.08$, $s = 0.05$, $z_\alpha = \pm 1.645$ at $\alpha = 0.05$ level of significance

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1.08 - 1.28}{0.05/\sqrt{40}} = -28.57$$

Since $z_{\text{cal}} (= -28.57) < z_{\alpha/2} = -1.64$, H_0 is rejected. Mean life of the boots is less than 1.28 and affected by use in the desert.

- 8.4** Let $H_0 : \mu = 8.9$ and $H_1 : \mu \neq 8.9$ (Two-tail test)

Given $n = 50$, $\bar{x} = 9.3$, $s = 1.8$, $z_{\alpha/2} = \pm 1.96$ at $\alpha = 0.05$ level of significance

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{9.3 - 8.9}{1.8/\sqrt{50}} = 1.574$$

Since $z_{\text{cal}} (= 1.574) < z_{\alpha/2} (= 1.96)$, H_0 is accepted, that is, claim is valid.

- 8.5** Let $H_0 : \mu = 40,000$ and $H_1 : \mu \neq 40,000$ (Two-tail test)

Given $n = 100$, $\bar{x} = 39,350$, $s = 3,260$, and $z_{\alpha/2} = \pm 2.58$ at $\alpha = 0.01$ level of significance

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{39,350 - 40,000}{3260/\sqrt{100}} = -1.994$$

Since $z_{\text{cal}} (= -1.994) > z_{\alpha/2} (= -2.58)$, H_0 is accepted. Thus the difference in the mean life of the tyres could be due to sampling error.

- 8.6** Let $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 > \mu_2$; μ_1 and μ_2 = mean height of Austrians and Englishmen, respectively.

Given, Austrian : $n_1 = 1600$, $\bar{x}_1 = 68.55$, $s_1 = 2.52$ and Englishmen; $n_2 = 6400$, $\bar{x}_2 = 67.85$, $s_2 = 2.56$

$$\begin{aligned} z &= \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{68.55 - 67.85}{\sqrt{\frac{(2.52)^2}{1600} + \frac{(2.56)^2}{6400}}} = 9.9 \end{aligned}$$

Since $z_{\text{cal}} = 9.9 > z_\alpha (= 2.58)$ for right tail test, H_0 is rejected. Austrian's are on the average taller than the Englishmen.

- 8.7** Let $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$; μ_1 and μ_2 = mean life of Philips and HMT electric bulbs, respectively

Given, Philips : $n_1 = 50$, $\bar{x}_1 = 1500$, $s_1 = 60$ and HMT: $n_2 = 50$, $\bar{x}_2 = 1512$, $s_2 = 80$

$$\begin{aligned} z &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{1500 - 1512}{\sqrt{\frac{(60)^2}{50} + \frac{(80)^2}{50}}} \\ &= -\frac{12}{14.14} = -0.848 \end{aligned}$$

Since $z_{\text{cal}} (= -0.848) > z_{\alpha/2} (= -2.58)$ at $\alpha = 0.01$ level of significance, H_0 is accepted. Mean life of the two makes is almost the same, difference (if any) is due to sampling error.

- 8.8** Given $n = 50$, $\bar{x} = 14.2$, $s = 5$, and $\alpha = 0.02$

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{14.2 - 15}{5/\sqrt{50}} = -1.13$$

Table value of $z = 1.13$ is 0.3708. Thus $p\text{-value} = 2(0.5000 - 0.3708) = 0.2584$. Since $p\text{-value} > \alpha$, H_0 is accepted.

- 8.9** Let $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 > \mu_2$; μ_1 and μ_2 = mean life of items produced by Method 1 and 2, respectively.

Given, Method 1: $n_1 = 64$, $\bar{x}_1 = 106$, $s_1 = 12$; Method 2: $n_2 = 64$, $\bar{x}_2 = 100$, $s_2 = 10$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{106 - 100}{\sqrt{\frac{(12)^2}{64} + \frac{(10)^2}{64}}} = 3.07$$

Since $z_{\text{cal}} (= 3.07) > z_\alpha (= 1.645)$ for a right-tailed test, H_0 is rejected. Method 1 is better than Method 2.

- 8.10** Let $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$; μ_1 and μ_2 = mean petrol mileage of two types of new cars, respectively

Given $n_1 = 36$, $\bar{x}_1 = 14$, $\sigma_1^2 = 1.5$ and $n_2 = 72$,

$\bar{x}_2 = 12.5$, $\sigma_2^2 = 2.0$

$$\begin{aligned} z &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{14 - 12.5}{\sqrt{\frac{1.5}{36} + \frac{2}{72}}} \\ &= \frac{1.5}{0.2623} = 5.703 \end{aligned}$$

Since $z_{\text{cal}} (= 5.703) > z_{\alpha/2} (= 2.58)$ at $\alpha = 0.01$ level of significance, H_0 is rejected. There is a significant difference in petrol consumption of the two types of new cars.

8.8 HYPOTHESIS TESTING FOR SINGLE POPULATION PROPORTION

Sometimes instead of testing a hypothesis pertaining to a population mean, a population proportion (a fraction, ratio or percentage) p of values that indicates the part of the population or sample having a particular attribute of interest is considered. For this, a random sample of size n is selected to compute the proportion of elements having a particular attribute of interest (also called success) in it as follows:

$$\bar{p} = \frac{\text{Number of successes in the sample}}{\text{Sample size}} = \frac{x}{n}$$

The value of this statistic is compared with a hypothesized population proportion p_0 so as to arrive at a conclusion about the hypothesis.

The three forms of null hypothesis and alternative hypothesis pertaining to the hypothesized population proportion p are as follows:

Null hypothesis	Alternative hypothesis
$H_0 : p = p_0$	$H_1 : p \neq p_0$ (Two-tailed test)
$H_0 : p \geq p_0$	$H_1 : p < p_0$ (Left-tailed test)
$H_0 : p \leq p_0$	$H_1 : p > p_0$ (Right-tailed test)

To conduct a test of a hypothesis, it is assumed that the sampling distribution of a proportion follows a standardized normal distribution. Then, using the value of the sample proportion \bar{p} and its standard deviation $\sigma_{\bar{p}}$, we compute a value for the z-test statistic as follows:

$$\text{Test statistic } z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

The comparison of the z-test statistic value to its critical (table) value at a given level of significance enables us to test the null hypothesis about a population proportion based on the difference between the sample proportion \bar{p} and the hypothesized population proportion.

Decision rule: Reject H_0 when

One-tailed test	Two-tailed test
<ul style="list-style-type: none"> $z_{\text{cal}} > z_\alpha$ or $z_{\text{cal}} < -z_\alpha$ when $H_1 : p < p_0$ 	<ul style="list-style-type: none"> $z_{\text{cal}} > z_{\alpha/2}$ or $z_{\text{cal}} < -z_{\alpha/2}$
$p\text{-value} < \alpha$	

8.8.1 Hypothesis Testing for Difference Between Two Population Proportions

Let two independent populations each having proportion and standard deviation of an attribute be as follows:

Population	Proportion	Standard Deviation
1	p_1	σ_{p_1}
2	p_2	σ_{p_2}

The hypothesis testing concepts developed in the previous section can be extended to test whether there is any difference between the proportions of these populations. The null hypothesis that there is no difference between two population proportions is stated as:

$$H_0 : p_1 = p_2 \text{ or } p_1 - p_2 = 0 \text{ and } H_1 : p_1 \neq p_2$$

The sampling distribution of difference in sample proportions $\bar{p}_1 - \bar{p}_2$ is based on the assumption that the difference between two population proportions, $p_1 - p_2$ is normally distributed. The standard deviation (or error) of sampling distribution of $\bar{p}_1 - \bar{p}_2$ is given by

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}; q_1 = 1 - p_1 \text{ and } q_2 = 1 - p_2$$

where the difference $\bar{p}_1 - \bar{p}_2$ between sample proportions of two independent simple random samples is the point estimator of the difference between two population proportions. Obviously expected value, $E(\bar{p}_1 - \bar{p}_2) = p_1 - p_2$.

Thus the z-test statistic for the difference between two population proportions is stated as:

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sigma_{\bar{p}_1 - \bar{p}_2}} = \frac{\bar{p}_1 - \bar{p}_2}{\sigma_{\bar{p}_1 - \bar{p}_2}}$$

Invariably, the standard error $\sigma_{\bar{p}_1 - \bar{p}_2}$ of difference between sample proportions is not known. Thus when a null hypothesis states that there is no difference between the population proportions, we combine two sample proportions \bar{p}_1 and \bar{p}_2 to get one unbiased estimate of population proportion as follows:

$$\text{Pooled estimate } \bar{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2}$$

The z-test statistic is then restated as:

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sigma_{\bar{p}_1 - \bar{p}_2}}; \sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\bar{p}(1-\bar{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

Example 8.12: An auditor claims that 10 per cent of customers' ledger accounts are carrying mistakes of posting and balancing. A random sample of 600 was taken to test the accuracy of posting and balancing and 45 mistakes were found. Are these sample results consistent with the claim of the auditor? Use 5 per cent level of significance.

Solution: Let us take the null hypothesis that the claim of the auditor is valid, that is,

$$H_0 : p = 0.10 \text{ and } H_1 : p \neq 0.10 \text{ (Two-tailed test)}$$

Given $\bar{p} = 45/600 = 0.075$, $n = 600$, and $\alpha = 5$ per cent. Thus using the z-test statistic

$$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} = \frac{0.075 - 0.10}{\sqrt{\frac{0.10 \times 0.90}{600}}} = -\frac{0.025}{0.0122} = -2.049$$

Since $z_{\text{cal}} (= -2.049)$ is less than its critical (table) value $z_\alpha (= -1.96)$ at $\alpha = 0.05$ level of significance, null hypothesis, H_0 is rejected. Hence, we conclude that the claim of the auditor is not valid.

Example 8.13: A manufacturer claims that at least 95 per cent of the equipments which he supplied to a factory conformed to the specification. An examination of the sample of 200 pieces of equipment revealed that 18 were faulty. Test the claim of the manufacturer.

Solution: Let us take the null hypothesis that at least 95 per cent of the equipments supplied conformed to the specification, that is,

$$H_0 : p \geq 0.95 \text{ and } H_1 : p < 0.95 \text{ (Left-tailed test)}$$

Given $\bar{p} =$ per cent of pieces conforming to the specification $= 1 - (18/100) = 0.91$ $n = 200$ and level of significance $\alpha = 0.05$. Thus using the z-test statistic,

$$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} = \frac{0.91 - 0.95}{\sqrt{\frac{0.95 \times 0.05}{200}}} = -\frac{0.04}{0.015} = -2.67$$

Since $z_{\text{cal}} (= -2.67)$ is less than its critical value $z_\alpha (= -1.645)$ at $\alpha = 0.05$ level of significance, the null hypothesis, H_0 is rejected. Hence we conclude that the proportion of equipments conforming to specifications is not 95 per cent.

Example 8.14: A company is considering two different television advertisements for promotion of a new product. Management believes that advertisement A is more effective than advertisement B. Two test market areas with virtually identical consumer characteristics

are selected: advertisement A is used in one area and advertisement B in the other area. In a random sample of 60 customers who saw advertisement A, 18 had tried the product. In a random sample of 100 customers who saw advertisement B, 22 had tried the product. Does this indicate that advertisement A is more effective than advertisement B, if a 5 per cent level of significance is used? [Delhi Univ., MFC 1996; MBA, 2000]

Solution: Let us take the null hypothesis that both advertisements are equally effective, that is,

$$H_0 : p_1 = p_2 \text{ and } H_1 : p_1 > p_2 \text{ (Right-tailed test)}$$

where p_1 and p_2 = proportion of customers who saw advertisement A and advertisement B respectively.

Given $n_1 = 60$, $\bar{p}_1 = 18/60 = 0.30$; $n_2 = 100$, $\bar{p}_2 = 22/100 = 0.22$ and level of significance $\alpha = 0.05$. Thus using the z-test statistic.

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}}; \quad p_1 = p_2$$

$$\begin{aligned} \text{where } s_{\bar{p}_1 - \bar{p}_2} &= \sqrt{\bar{p}(1-\bar{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}; \quad q = 1 - p \\ &= \sqrt{0.25 \times 0.75 \left(\frac{1}{60} + \frac{1}{100} \right)} = \sqrt{0.1875 \left(\frac{160}{600} \right)} = 0.0707; \\ \bar{p} &= \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} = \frac{60(18/60) + 100(22/100)}{60 + 100} \\ &= \frac{18 + 22}{160} = \frac{40}{160} = 0.25 \end{aligned}$$

Substituting values in z-test statistic, we have

$$z = \frac{0.30 - 0.22}{0.0707} = \frac{0.08}{0.0707} = 1.131$$

Since $z_{\text{cal}} = 1.131$ is less than its critical value $z_\alpha = 1.645$ at $\alpha = 0.05$ level of significance, the null hypothesis, H_0 is accepted. Hence we conclude that there is no significant difference in the effectiveness of the two advertisements.

Example 8.15: In a simple random sample of 600 men taken from a big city, 400 are found to be smokers. In another simple random sample of 900 men taken from another city 450 are smokers. Do the data indicate that there is a significant difference in the habit of smoking in the two cities? [Raj Univ., MCom, 1998; Punjab Univ., MCom, 1996]

Solution: Let us take the null hypothesis that there is no significant difference in the habit of smoking in the two cities, that is,

$$H_0 : p_1 = p_2 \text{ and } H_1 : p_1 \neq p_2 \text{ (Two-tailed test)}$$

where p_1 and p_2 = proportion of men found to be smokers in the two cities.

Given, $n_1 = 600$, $\bar{p}_1 = 400/600 = 0.667$; $n_2 = 900$, $\bar{p}_2 = 450/900 = 0.50$ and level of significance $\alpha = 0.05$. Thus using the z-test statistic

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}}; \quad p_1 = p_2$$

$$\begin{aligned} \text{where } s_{\bar{p}_1 - \bar{p}_2} &= \sqrt{\bar{p}(1-\bar{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}; \quad q = 1 - p \\ &= \sqrt{0.567 \times 0.433 \left(\frac{1}{600} + \frac{1}{900} \right)} = \sqrt{0.245 (0.002)} = 0.026; \\ \bar{p} &= \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} = \frac{600 (400/600) + 900 (450/900)}{600 + 900} \\ &= \frac{400 + 450}{1500} = \frac{850}{1500} = 0.567 \end{aligned}$$

Substituting values in z-test statistic, we have

$$z = \frac{0.667 - 0.500}{0.026} = \frac{0.167}{0.026} = 6.423$$

Since $z_{\text{cal}} = 6.423$ is greater than its critical value $z_{\alpha/2} = 2.58$, at $\alpha/2 = 0.025$ level of significance, the null hypothesis, H_0 is rejected. Hence we conclude that there is a significant difference in the habit of smoking in two cities.

8.9 HYPOTHESIS TESTING FOR A BINOMIAL PROPORTION

The sampling of traits or attributes is considered as drawing of samples from a population whose elements have a particular trait of interest. For example, in the study of attribute such 'good or acceptable' pieces of items manufactured by company, a sample of suitable size may be taken from a given lot of items to classify them as good or not.

Instead of examining the hypothesis regarding *proportion of elements having the same trait (called success)* in a sample as discussed in the previous section, we could examine the *number of successes* in a sample. The z-test statistic for determining the magnitude of the difference between the number of successes in a sample and the hypothesized (expected) number of successes in the population is given by

$$z = \frac{\text{Sample estimate} - \text{Expected value}}{\text{Standard error of estimate}} = \frac{x - np}{\sqrt{npq}}$$

Recalling from previous discussion that although sampling distribution of the number of successes in the sample follows a binomial distribution having its mean $\mu = np$ and standard deviation \sqrt{npq} , the normal distribution provides a good approximation to the binomial distribution provided the sample size is large, that is, both $np \geq 5$ and $n(1-p) \geq 5$.

Example 8.16: Suppose the production manager implements a newly developed sealing system for boxes. He takes a random sample of 200 boxes from the daily output and finds that 12 need rework. He is interested to determine whether the new sealing system has increased defective packages below 10 per cent. Use 1 per cent level of significance

Solution: Let us state the null and alternative hypotheses as follows:

$$H_0 : p \geq 0.10 \quad \text{and} \quad H_1 : p < 0.10 \quad (\text{Left-tailed test})$$

Given $n = 200$, $p = 0.10$, and level of significance $\alpha = 0.01$. Applying the z-test statistic

$$z = \frac{x - np}{\sqrt{npq}} = \frac{12 - 200(0.10)}{\sqrt{200(0.10)(0.90)}} = \frac{12 - 20}{\sqrt{18}} = -1.885$$

Since $z_{\text{cal}} (-1.885)$ is more than its critical value $z_{\alpha} = -2.33$ for one-tailed test at $\alpha = 0.01$ level of significance, the null hypothesis, H_0 is accepted. Hence we conclude that the proportion of defective packages with the new sealing system is more than 10 per cent.

Example 8.17: In 324 throws of a six-faced dice, odd points appeared 180 times. Would you say that the dice is fair at 5 per cent level of significance?

[MD Univ., MCom, 1997]

Solution: Let us take the hypothesis that the dice is fair, that is,

$$H_0 : p = 162/324 = 0.5 \quad \text{and} \quad H_1 : p \neq 0.5 \quad (\text{Two-tailed test})$$

Given $n = 324$, $p = q = 0.5$ (i.e., 162 odd or even points out of 324 throws). Applying the z-test statistic:

$$z = \frac{x - np}{\sqrt{npq}} = \frac{180 - 162}{\sqrt{324 \times 0.5 \times 0.5}} = \frac{18}{9} = 2$$

Since $z_{\text{cal}} = 2$ is more than its critical value $z_{\alpha/2} = 1.96$ at $\alpha/2 = 0.025$ significance level, the null hypothesis H_0 is rejected. Hence we conclude that the dice is not fair.

Example 8.18: Of those women who are diagnosed to have early-stage breast cancer, one-third eventually die of the disease. Suppose an NGO launch a screening programme to provide for the early detection of breast cancer and to increase the survival rate of those diagnosed to have the disease. A random sample of 200 women was selected from among those who were periodically screened and who were diagnosed to have the disease. If 164 women in the sample of 200 survive the disease, can screening programme be considered effective? Test using $\alpha=0.01$ level of significance and explain the conclusions from your test.

Solution: Let us take the null hypothesis that the screening programme was effective, that is,

$$H_0: p = 1 - (1/3) = 2/3 \text{ and } H_1: p > 2/3$$

Given $n = 200$, $p = 2/3$, $q = 1/3$ and $\alpha = 0.05$. Applying the z-test statistic,

$$\begin{aligned} z &= \frac{x - np}{\sqrt{npq}} = \frac{164 - 200 \times (2/3)}{\sqrt{200 \times (2/3)(1/3)}} = \frac{164 - 133.34}{\sqrt{44.45}} \\ &= \frac{30.66}{6.66} = 4.60 \end{aligned}$$

Since $z_{\text{cal}} = 4.60$ is greater than its critical value $z_\alpha = 2.33$ at $\alpha = 0.01$ significance level, the null hypothesis, H_0 is rejected. Hence we conclude that the screening programme was not effective.

Self-Practice Problems 8B

- 8.11** A company manufacturing a certain type of breakfast cereal claims that 60 per cent of all housewives prefer that type to any other. A random sample of 300 housewives contains 165 who do prefer that type. At 5 per cent level of significance, test the claim of the company.
- 8.12** An auditor claims that 10 per cent of a company's invoices are incorrect. To test this claim a random sample of 200 invoices is checked and 24 are found to be incorrect. At 1 per cent significance level, test whether the auditor's claim is supported by the sample evidence.
- 8.13** A sales clerk in the department store claims that 60 per cent of the shoppers entering the store leave without making a purchase. A random sample of 50 shoppers showed that 35 of them left without buying anything. Are these sample results consistent with the claim of the sales clerk? Use a significance level of 0.05.
[Delhi Univ., MBA, 1998, 2001]
- 8.14** A dice is thrown 49,152 times and of these 25,145 yielded either 4, 5, or 6. Is this consistent with the hypothesis that the dice must be unbiased?
- 8.15** A coin is tossed 100 times under identical conditions independently yielding 30 heads and 70 tails. Test at 1 per cent level of significance whether or not the coin is unbiased. State clearly the null hypothesis and the alternative hypothesis.
- 8.16** Before an increase in excise duty on tea, 400 people out of a sample of 500 persons were found to be tea drinkers. After an increase in the duty, 400 persons were known to be tea drinkers in a sample of 600 people. Do you think that there has been a significant decrease in the consumption of tea after the increase in the excise duty?
[Delhi Univ., MCom, 1998; MBA, 2000]
- 8.17** In a random sample of 1000 persons from UP 510 were found to be consumers of cigarettes. In another sample of 800 persons from Rajasthan, 480 were found to be consumers of cigarettes. Do the data reveal a significant difference between UP and Rajasthan so far as the proportion of consumers of cigarettes in concerned?
[MC Univ., M.Com, 1996]
- 8.18** In a random sample of 500 persons belonging to urban areas, 200 are found to be using public transport. In another sample of 400 persons belonging to rural area 200 area found to be using public transport. Do the data reveal a significant difference between urban and rural areas so far as the proportion of commuters of public transport is concerned (use 1 per cent level of significance).
[Bharathidasan Univ., MCom, 1998]
- 8.19** A machine puts out 10 defective units in a sample of 200 units. After the machine is overhauled it puts out 4 defective units in a sample of 100 units. Has the machine been improved?
[Madras Univ., MCom, 1996]
- 8.20** 500 units from a factory are inspected and 12 are found to be defective, 800 units from another factory are inspected and 12 are found to be defective. Can it be concluded that at 5 per cent level of significance production at the second factory is better than in first factory?
[Kurukshetra Univ., MBA, 1996; Delhi Univ., MBA, 2002]
- 8.21** In a hospital 480 female and 520 male babies were born in a week. Do these figures confirm the hypothesis that females and males are born in equal number?
[Madras Univ., MCom, 1997]
- 8.22** A wholesaler of eggs claims that only 4 per cent of the eggs supplied by him are defective. A random sample of 600 eggs contained 36 defectives. Test the claim of the wholesaler.
[IGNOU, 1997]

Hints and Answers

- 8.11** Let $H_0 : p = 60$ per cent and $H_1 : p < 60$ per cent (One tailed test)

Given, sample proportion, $\bar{p} = 165/300 = 0.55$; $n = 300$ and $z_{\alpha} = 1.645$ at $\alpha = 5$ per cent

$$z = \frac{\bar{p} - p_0}{\sqrt{\frac{pq}{n}}} = \frac{0.55 - 0.60}{\sqrt{\frac{0.60 \times 0.40}{300}}} = -1.77$$

Since $z_{\text{cal}} (= -1.77)$ is less than its critical value $z_{\alpha} = -1.645$, the H_0 is rejected. Percentage preferring the breakfast cereal is lower than 60 per cent.

- 8.12** Let $H_0 : p = 10$ per cent and $H_1 : p \neq 10$ per cent (Two-tailed test)

Given, sample proportion, $\bar{p} = 24/200 = 0.12$; $n = 200$ and $z_{\alpha/2} = 2.58$ at $\alpha = 1$ per cent

$$z = \frac{\bar{p} - p_0}{\sqrt{\frac{pq}{n}}} = \frac{0.12 - 0.10}{\sqrt{\frac{0.10 \times 0.90}{200}}} = 0.943$$

Since $z_{\text{cal}} (= 0.943)$ is less than its critical value $z_{\alpha/2} = 2.58$, the H_0 is accepted. Thus the percentage of incorrect invoices is consistent with the auditor's claim of 10 per cent.

- 8.13** Let $H_0 : p = 60$ per cent and $H_1 : p \neq 60$ per cent (Two-tailed test)

Given, sample proportion, $\bar{p} = 35/60 = 0.70$; $n = 50$ and $z_{\alpha/2} = 1.96$ at $\alpha = 5$ per cent

$$z = \frac{\bar{p} - p_0}{\sqrt{\frac{pq}{n}}} = \frac{0.70 - 0.60}{\sqrt{\frac{0.60 \times 0.40}{50}}} = 1.44$$

Since $z_{\text{cal}} (= 1.44)$ is less than its critical value $z_{\alpha/2} = 1.96$, the H_0 is accepted. Claim of the sales clerk is valid.

- 8.14** Let $H_0 : p = 50$ per cent and $H_1 : p \neq 50$ per cent (Two-tailed test)

Given, sample proportion of success $p = 25,145/49,152 = 0.512$; $n = 49,152$

$$z = \frac{\bar{p} - p_0}{\sqrt{\frac{pq}{n}}} = \frac{0.512 - 0.50}{\sqrt{\frac{0.50 \times 0.50}{49,152}}} = \frac{0.012}{0.002} = 6.0$$

Since $z_{\text{cal}} (= 6.0)$ is more than its critical value $z_{\alpha/2} = 2.58$ at $\alpha = 0.01$, the H_0 is rejected. Dice is biased.

- 8.15** Let $H_0 : p = 50$ per cent and $H_1 : p \neq 50$ per cent (Two-tailed test)

Given, $n = 100$, sample proportion of success

$\bar{p} = 30/100 = 0.30$ and $z_{\alpha/2} = 2.58$ at $\alpha = 0.01$

$$z = \frac{\bar{p} - p_0}{\sqrt{\frac{pq}{n}}} = \frac{0.30 - 0.50}{\sqrt{\frac{0.50 \times 0.50}{100}}} = -\frac{0.20}{0.05} = -4$$

Since $z_{\text{cal}} (= -4)$ is less than its critical value $z_{\alpha/2} = -2.58$, the H_0 is rejected.

- 8.16** Let $H_0 : p = 400/500 = 0.80$ and $H_1 : p < 0.80$ (One-tailed test)

Given $n_1 = 500$, $n_2 = 600$, $\bar{p}_1 = 400/500 = 0.80$, $\bar{p}_2 = 400/600 = 0.667$ and $z_{\alpha} = 2.33$ at $\alpha = 0.01$

$$\bar{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} = \frac{400 + 400}{500 + 600} = 0.727;$$

$$q = 1 - 0.727 = 0.273$$

$$\begin{aligned}s_{\bar{p}_1 - \bar{p}_2} &= \sqrt{\bar{p} \bar{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \\ &= \sqrt{0.727 \times 0.273 \left(\frac{1}{500} + \frac{1}{600} \right)} = 0.027\end{aligned}$$

$$z = \frac{\bar{p}_1 - \bar{p}_2}{s_{\bar{p}_1 - \bar{p}_2}} = \frac{0.80 - 0.667}{0.027} = 4.93$$

Since $z_{\text{cal}} (= 4.93)$ is more than its critical value $z_{\alpha} = 2.33$, the H_0 is rejected. Decrease in the consumption of tea after the increase in the excise duty is significant.

- 8.17** Let $H_0 : p_1 = p_2$ and $H_1 : p_1 \neq p_2$ (Two-tailed test)

Given, UP: $n_1 = 1000$, $\bar{p}_1 = 510/1000 = 0.51$; Rajasthan: $n_2 = 800$, $\bar{p}_2 = 480/800 = 0.60$

$$\bar{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} = \frac{510 + 480}{1000 + 800} = 0.55;$$

$$q = 1 - 0.55 = 0.45$$

$$\begin{aligned}s_{\bar{p}_1 - \bar{p}_2} &= \sqrt{\bar{p} \bar{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \\ &= \sqrt{0.55 \times 0.45 \left(\frac{1}{1000} + \frac{1}{800} \right)} = 0.024.\end{aligned}$$

$$z = \frac{\bar{p}_1 - \bar{p}_2}{s_{\bar{p}_1 - \bar{p}_2}} = \frac{0.51 - 0.60}{0.024} = -3.75$$

Since $z_{\text{cal}} (= -3.75)$ is less than its critical value $z_{\alpha/2} = -2.58$, the H_0 is rejected. The proportion of consumers of cigarettes in the two states is significant.

- 8.18** Let $H_0 : p_1 = p_2$ and $H_1 : p_1 \neq p_2$ (Two-tailed test)

Given, Urban area: $n_1 = 500$, $\bar{p}_1 = 200/500 = 0.40$;

Rural area: $n_2 = 200$, $\bar{p}_2 = 200/400 = 0.50$

$$\bar{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} = \frac{200 + 200}{500 + 400} = 0.44;$$

$$q = 1 - \bar{p} = 0.55$$

$$\begin{aligned}s_{\bar{p}_1 - \bar{p}_2} &= \sqrt{\bar{p} \bar{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \\ &= \sqrt{0.44 \times 0.55 \left(\frac{1}{500} + \frac{1}{400} \right)} = 0.033\end{aligned}$$

$$z = \frac{\bar{p}_1 - \bar{p}_2}{s_{\bar{p}_1 - \bar{p}_2}} = \frac{0.40 - 0.50}{0.033} = -3.03$$

Since $z_{\text{cal}} = -3.03$ is less than its critical value $z_{\alpha/2} = -2.58$, the H_0 is rejected. Proportion of commuters of public transport in urban and rural areas is significant.

8.19 Let $H_0 : p_1 \leq p_2$ and $H_1 : p_1 > p_2$ (One-tailed test)

Given, Before overhaul: $n_1 = 200$, $\bar{p}_1 = 10/200 = 0.05$;

After overhaul: $n_2 = 100$, $\bar{p}_2 = 4/100 = 0.04$

$$\bar{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} = \frac{10 + 4}{200 + 100} = 0.047;$$

$$q = 1 - p = 0.953$$

$$s_{\bar{p}_1 - \bar{p}_2} = \sqrt{0.047 \times 0.953 \left(\frac{1}{200} + \frac{1}{100} \right)} = 0.026;$$

$$z = \frac{\bar{p}_1 - \bar{p}_2}{s_{\bar{p}_1 - \bar{p}_2}} = \frac{0.05 - 0.04}{0.026} = 0.385$$

Since z_{cal} ($= 0.385$) is less than its critical value $z_\alpha = 1.645$ at $\alpha = 0.05$, the H_0 is accepted.

8.20 Let $H_0 : p_1 \leq p_2$ and $H_1 : p_1 > p_2$ (One-tailed test)

Given $n_1 = 500$, $\bar{p}_1 = 12/500 = 0.024$, $n_2 = 800$,

$$\bar{p}_2 = 12/800 = 0.015$$

$$\bar{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} = \frac{12 + 12}{500 + 800} = 0.018;$$

$$q = 1 - p = 0.982$$

$$s_{\bar{p}_1 - \bar{p}_2} = \sqrt{\bar{p} \bar{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$= \sqrt{0.018 \times 0.982 \left(\frac{1}{500} + \frac{1}{800} \right)} = 0.0076$$

$$z = \frac{\bar{p}_1 - \bar{p}_2}{s_{\bar{p}_1 - \bar{p}_2}} = \frac{0.024 - 0.015}{0.0076} = 1.184$$

Since $z_{\text{cal}} = 1.184$ is less than its critical value $z_\alpha = 1.645$ at $\alpha = 0.05$, the H_0 is accepted. Production in second factory is better than in the first factory.

8.21 Let $H_0 : p_1 = p_2$ and $H_1 : p_1 \neq p_2$ (Two-tailed test)

Given $n = 480 + 520 = 1000$, $p = q = 0.5$.

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sigma_{\bar{p}_1 - \bar{p}_2}} = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{npq}} = \frac{520 - 480}{\sqrt{1000(0.5)(0.5)}} \\ = \frac{40}{15.81} = 2.53$$

Since, $z_{\text{cal}} = 2.53$ is greater than its critical value $z_{\alpha/2} = 1.96$ at $\alpha/2 = 0.025$, the H_0 is rejected.

8.22 Let $H_0 : p = 4$ per cent and $H_1 : p \neq 4$ per cent (Two-tailed test)

Given $n = 600$, $\bar{p} = 36/600 = 0.06$

$$\text{Confidence limits: } \bar{p} \pm z_\alpha \sqrt{\frac{pq}{n}}$$

$$= 0.06 \pm 1.96 \sqrt{(0.04 \times 0.96)/600} \\ = 0.06 \pm 0.016; \quad 0.44 \leq p \leq 0.076$$

Since probability 0.04 of the claim does not fall into the confidence limit, the claim is rejected.

8.10 HYPOTHESIS TESTING FOR POPULATION MEAN WITH SMALL SAMPLES

When the sample size is small (i.e., less than 30), the central limit theorem does not assure us to assume that the sampling distribution of a statistic such as mean \bar{x} , proportion \bar{p} , is normal. Consequently when testing a hypothesis with small samples, we must assume that the samples come from a normally or approximately normally distributed population. Under these conditions, the sampling distribution of sample statistic such as \bar{x} and \bar{p} is normal but the critical values of \bar{x} or \bar{p} depend on whether or not the population standard deviation σ is known. When the value of the population standard deviation σ is not known, its value is estimated by computing the standard deviation of sample s and the standard error of the mean is calculated by using the formula, $\sigma_{\bar{x}} = s/\sqrt{n}$. When we do this, the resulting sampling distribution may not be normal even if sampling is done from a normally distributed population. In all such cases the sampling distribution turns out to be the *Student's t-distribution*.

Sir William Gosset of Ireland in early 1900, under his pen name 'Student', developed a method for hypothesis testing popularly known as the '**t-test**'. It is said that Gosset was employed by Guinness Breway in Dublin, Ireland which did not permit him to publish his research findings under his own name, so he published his research findings in 1905 under the pen name 'Student'.

t-test: A hypothesis test for comparing two independent population means using the means of two small samples.

8.10.1 Properties of Student's t-Distribution

If small samples of size n (≤ 30) are drawn from normal population with mean μ and for each sample we compute the sample statistic of interest, then probability density function of the *t-distribution* with degrees of freedom v (a Greek letter nu) is given by

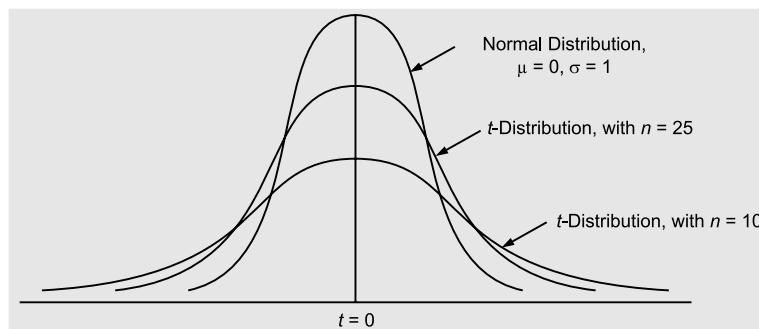
$$y = \frac{y_0}{\left(1 + \frac{t^2}{n-1}\right)^{\frac{n}{2}}} = \frac{y_0}{\left(1 + \frac{t^2}{v}\right)^{\frac{(v+1)}{2}}}; -\infty \leq t \leq \infty$$

where y_0 is a constant depending on sample size n such that the total area under the curve is unity and $v = n - 1$.

- (i) As t appears in even power in probability density function, the t -distribution is symmetrical about the line $t = 0$ like the normal distribution.
- (ii) The shape of the t -distribution depends on the sample size n . As n increases, the variability of t decreases. In other words, for large values of degrees of freedom, the t -distribution tends to a standard normal distribution. This implies that for different degrees of freedom, the shape of the t -distribution also differs, as shown in Fig. 8.9. Eventually when n is infinitely large, the t and z distributions are identical.
- (iii) The t -distribution is less peaked than normal distribution at the centre and higher in the tails.
- (iv) The t -distribution has greater dispersion than standard normal distribution with heavier tails, i.e. the t -curve does not approach x -axis as quickly as z does. This is because the t -statistic involves two random variables \bar{x} and s , where as z -statistic involves only the sample mean \bar{x} . The variance of t -distribution is defined only when $v \geq 3$ and is given by $\text{var}(t) = v/(v - 2)$.

Figure 8.9

Comparison of t -Distribution with Standard Normal Distribution



- (v) The value of y attains its maximum value at $t = 0$ so that the mode coincides with the mean. The limiting value of t -distribution when $v \rightarrow \infty$ is given by $y = y_0 e^{-t^2/2}$. It follows that t is normally distributed for a large sample size,
- (vi) The degrees of freedom refers to the number of independent squared deviations in s^2 that are available for estimating σ^2 .

Uses of t -Distribution

There are various uses of t -distribution. A few of them are as follows:

- (i) Hypothesis testing for the population mean.
- (ii) Hypothesis testing for the difference between two populations means with independent samples.
- (iii) Hypothesis testing for the difference between two populations means with dependent samples.
- (iv) Hypothesis testing for an observed coefficient of correlation including partial and rank correlations.
- (v) Hypothesis testing for an observed regression coefficient.

8.10.2 Hypothesis Testing for Single Population Mean

The test statistic for determining the difference between the sample mean \bar{x} and population mean μ is given by

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}}; s = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}}$$

where s is an unbiased estimation of unknown population standard deviation σ . This test statistic has a t -distribution with $n - 1$ degrees of freedom.

Confidence Interval The confidence interval estimate of the population mean μ when unknown population standard deviation σ is estimated by sample standard deviation s , is given by:

- Two-tailed test : $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$; α = level of significance
- One-tailed test : $\bar{x} \pm t_{\alpha} \frac{s}{\sqrt{n}}$; α = level of significance

where t -test statistic value is based on a t -distribution with $n - 1$ degrees of freedom and $1 - \alpha$ is the confidence coefficient.

- Null hypothesis, $H_0: \mu = \mu_0$
- Alternative hypothesis:

One-tailed test	Two-tailed test
$H_1: \mu > \mu_0$ or $\mu < \mu_0$	$H_1: \mu \neq \mu_0$

Decision Rule: Rejected H_0 at the given degrees of freedom $n-1$ and level of significance when

One-tailed test	Two-tailed test
• $t_{\text{cal}} > t_{\alpha}$ or $t_{\text{cal}} < -t_{\alpha}$ for $H_1: \mu < \mu_0$	$t_{\text{cal}} > t_{\alpha/2}$ or $t_{\text{cal}} < -t_{\alpha/2}$
• Reject H_0 when $p\text{-value} < \alpha$	

Example 8.19: The average breaking strength of steel rods is specified to be 18.5 thousand kg. For this a sample of 14 rods was tested. The mean and standard deviation obtained were 17.85 and 1.955, respectively. Test the significance of the deviation.

Solution: Let us take the null hypothesis that there is no significant deviation in the breaking strength of the rods, that is,

$$H_0: \mu = 18.5 \quad \text{and} \quad H_1: \mu \neq 18.5 \quad (\text{Two-tailed test})$$

Given, $n = 14$, $\bar{x} = 17.85$, $s = 1.955$, $df = n - 1 = 13$, and $\alpha = 0.05$ level of significance. The critical value of t at $df = 13$ and $\alpha/2 = 0.025$ is $t_{\alpha/2} = 2.16$.

Using the z-test statistic,

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{17.85 - 18.5}{\frac{1.955}{\sqrt{14}}} = -\frac{0.65}{0.522} = -1.24$$

Since $t_{\text{cal}} (= -1.24)$ value is more than its critical value $t_{\alpha/2} = -2.16$ at $\alpha/2 = 0.025$ and $df = 13$, the null hypothesis H_0 is accepted. Hence we conclude that there is no significant deviation of sample mean from the population mean.

Example 8.20: An automobile tyre manufacturer claims that the average life of a particular grade of tyre is more than 20,000 km when used under normal conditions. A random sample of 16 tyres was tested and a mean and standard deviation of 22,000 km and 5000 km, respectively were computed. Assuming the life of the tyres in km to be approximately normally distributed, decide whether the manufacturer's claim is valid.

Solution: Let us take the null hypothesis that the manufacturer's claim is valid, that is,

$$H_0: \mu \geq 20,000 \quad \text{and} \quad H_1: \mu < 20,000 \quad (\text{Left-tailed test})$$

Given, $n = 16$, $\bar{x} = 22,000$, $s = 5000$, $df = 15$ and $\alpha = 0.05$ level of significance. The critical value of t at $df = 15$ and $\alpha = 0.05$ is $t_{\alpha} = 1.753$. Using the z-test statistic,

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{22,000 - 20,000}{5000/\sqrt{16}} = \frac{2000}{1250} = 1.60$$

Since t_{cal} ($= 1.60$) value is less than its critical value $t_{\alpha/2} = 1.753$, $\alpha = 0.05$ and $df = 15$ at the null hypothesis H_0 is accepted. Hence we conclude that the manufacturer's claim is valid.

Example 8.21: A fertilizer mixing machine is set to give 12 kg of nitrate for every 100 kg of fertilizer. Ten bags of 100 kg each are examined. The percentage of nitrate so obtained is: 11, 14, 13, 12, 13, 12, 13, 14, 11, and 12. Is there reason to believe that the machine is defective?

Solution: Let us take the null hypothesis that the machine produces 12 kg of nitrate for every 100 kg of fertilizer, and is not defective, that is,

$$H_0 : \mu = 12 \quad \text{and} \quad H_1 : \mu \neq 12 \text{ (Two-tailed test)}$$

Given $n = 10$, $df = 9$, and $\alpha = 0.05$, critical value $t_{\alpha/2} = 2.262$ at $df = 9$ and $\alpha/2 = 0.025$.

Assuming that the weight of nitrate in bags is normally distributed and its standard deviation is unknown. The sample mean \bar{x} and standard deviation s values are calculated as shown in Table 8.4.

Table 8.4: Calculations of Sample Mean \bar{x} and Standard Deviation s

Variable x	Deviation, $d = x - 12$	d^2
11	-1	1
14	2	4
13	1	1
12	0	0
13	1	1
12	0	0
13	1	1
14	2	4
11	-1	1
12	0	0
125	5	13

$$\begin{aligned}\bar{x} &= \frac{\sum x}{n} = \frac{125}{10} = 12.5 \quad \text{and} \quad s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}} \\ &= \sqrt{\frac{13}{9} - \frac{(5)^2}{10(9)}} = 1.08\end{aligned}$$

Using the z -test statistic, we have

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{12.5 - 12}{\frac{1.08}{\sqrt{10}}} = \frac{0.50}{0.341} = 1.466$$

Since t_{cal} ($= 1.466$) value is less than its critical value $t_{\alpha/2} = 2.262$, at $\alpha/2 = 0.025$ and $df = 9$, the null hypothesis H_0 is accepted. Hence we conclude that the manufacturer's claim is valid, that is, the machine is not defective.

Example 8.22: A random sample of size 16 has the sample mean 53. The sum of the squares of deviation taken from the mean value is 150. Can this sample be regarded as taken from the population having 56 as its mean? Obtain 95 per cent and 99 per cent confidence limits of the sample mean.

Solution: Let us take the null hypothesis that the population mean is 56, i.e.

$$H_0 : \mu = 56 \quad \text{and} \quad H_1 : \mu \neq 56 \text{ (Two-tailed test)}$$

$$\text{Given, } n = 16, df = n - 1 = 15, \bar{x} = 53; s = \sqrt{\frac{\sum (x - \bar{x})^2}{(n-1)}} = \sqrt{\frac{150}{15}} = 3.162$$

- 95 per cent confidence limit

$$\bar{x} \pm t_{0.05} \frac{s}{\sqrt{n}} = 53 \pm 2.13 \frac{3.162}{\sqrt{16}} = 53 \pm 2.13 (0.790) = 53 \pm 1.683$$

- 99 per cent confidence limit

$$\bar{x} \pm t_{0.01} \frac{s}{\sqrt{n}} = 53 \pm 2.95 \frac{3.162}{\sqrt{16}} = 53 \pm 2.33$$

8.10.3 Hypothesis Testing for Difference of Two Population Means (Independent Samples)

For comparing the mean values of two normally distributed populations, we draw independent random samples of sizes n_1 and n_2 from the two populations. If μ_1 and μ_2 are the mean values of two populations, then our aim is to estimate the value of the difference $\mu_1 - \mu_2$ between mean values of the two populations.

Since sample mean \bar{x}_1 and \bar{x}_2 are the best point estimators to draw inferences regarding μ_1 and μ_2 respectively, therefore the difference between the sample means of the two independent simple random samples, $\bar{x}_1 - \bar{x}_2$, is the best point estimator of the difference $\mu_1 - \mu_2$.

The sampling distribution of $\bar{x}_1 - \bar{x}_2$ has the following properties:

- Expected value : $E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2$

This implies that the sample statistic $(\bar{x}_1 - \bar{x}_2)$ is an unbiased point estimator of $\mu_1 - \mu_2$.

- Variance : $\text{Var}(\bar{x}_1 - \bar{x}_2) = \text{Var}(\bar{x}_1) + \text{Var}(\bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

If the population standard deviations σ_1 and σ_2 are known, then the large sample interval estimation can also be used for the small sample case. But if these are unknown, then these are estimated by the sample standard deviations s_1 and s_2 . It is needed if sampling distribution is not normal even if sampling is done from two normal populations. The logic for this is the same as that for a single population case. Thus *t*-distribution is used to develop a small sample interval estimate for $\mu_1 - \mu_2$.

Population Variances are Unknown But Equal

If population variances σ_1^2 and σ_2^2 are unknown but equal, that is, both populations have exactly the same shape and $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then standard error of the difference in two sample means $\bar{x}_1 - \bar{x}_2$ can be written as:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

In such a case we need not estimate σ_1^2 and σ_2^2 separately and therefore data from two samples can be combined to get a pooled, single estimate of σ^2 . If we use the sample estimate s^2 for the population variance σ^2 , then the pooled variance estimator of σ^2 is given by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

This single variance estimator s^2 is a *weighted average* of the values of s_1^2 and s_2^2 in which weights are based on the degrees of freedom $n_1 - 1$ and $n_2 - 1$. Thus the point estimate of $\sigma_{\bar{x}_1 - \bar{x}_2}$ when $\sigma_1^2 = \sigma_2^2 = \sigma^2$ is given by

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Since $s_1 = \sqrt{\frac{\sum (x_1 - \bar{x}_1)^2}{(n_1 - 1)}}$ and $s_2 = \sqrt{\frac{\sum (x_2 - \bar{x}_2)^2}{(n_2 - 1)}}$, therefore the pooled variance s^2 can also be calculated as

$$s^2 = \frac{\sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

Following the same logic as discussed earlier, the t -test statistic is defined as

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

The sampling distribution of this t -statistic is approximated by the t -distribution with $n_1 + n_2 - 2$ degrees of freedom.

Null hypothesis: $H_0: \mu_1 - \mu_2 = d_0$

Alternative hypothesis:

One-tailed Test

$$H_1: (\mu_1 - \mu_2) > d_0 \text{ or } (\mu_1 - \mu_2) < d_0$$

Two-tailed Test

$$H_1: \mu_1 - \mu_2 \neq d_0$$

Decision Rule Rejected H_0 at $df = n_1 + n_2 - 2$ and at specified level of significance α when

One-tailed test	Two-tailed test
$t_{\text{cal}} > t_\alpha$ or $t_{\text{cal}} < -t_\alpha$ for $H_1: (\mu_1 - \mu_2) < d_0$	$t > t_{\alpha/2}$ or $t_{\text{cal}} < -t_{\alpha/2}$

Confidence Interval: The confidence interval estimate of the difference between populations means for small samples of size $n_1 < 30$ and/or $n_2 < 30$ with unknown σ_1 and σ_2 estimated by s_1 and s_2 is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_{\bar{x}_1 - \bar{x}_2}$$

where $t_{\alpha/2}$ is the critical value of t . The value of $t_{\alpha/2}$ depends on the t -distribution with $n_1 + n_2 - 2$ degrees of freedom and confidence coefficient $1 - \alpha$.

Population Variances are Unknown and Unequal

When two population variances are not equal, we may estimate the standard error $\sigma_{\bar{x}_1 - \bar{x}_2}$ of the statistic $(\bar{x}_1 - \bar{x}_2)$ by sample variances s_1^2 and s_2^2 in place of σ_1^2 and σ_2^2 . Thus an estimate of standard error of $\bar{x}_1 - \bar{x}_2$ is given by

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The sampling distribution of a t -test statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

is approximated by t -distribution with degrees of freedom given by

$$\text{Degrees of freedom (df)} = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

The number of degrees of freedom in this case is less than that obtained in Case 1 above.

Example 8.23: In a test given to two groups of students, the marks obtained are as follows:

First group : 18 20 36 50 49 36 34 49 41
 Second group : 29 28 26 35 30 44 46

Examine the significance of the difference between the arithmetic mean of the marks secured by the students of the above two groups.

[Madras Univ., MCom, 1997; MD Univ., MCom, 1998]

Solution: Let us take the null hypothesis that there is no significant difference in arithmetic mean of the marks secured by students of the two groups, that is,

$$H_0: \mu_1 - \mu_2 = 0 \text{ or } \mu_1 = \mu_2 \text{ and } H_1: \mu_1 \neq \mu_2 \text{ (Two-tailed test)}$$

Since sample size in both the cases is small and sample variances are not known, apply *t*-test statistic to test the null hypothesis

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Calculations of sample means \bar{x}_1 , \bar{x}_2 and pooled sample standard deviations are shown in Table 8.5.

Table 8.5: Calculation for \bar{x}_1 , \bar{x}_2 and s

First Group x_1	$x_1 - \bar{x}_1$ = $x_1 - 37$	$(x_1 - \bar{x}_1)^2$	x_2	$x_2 - \bar{x}_2$ = $x_2 - 34$	$(x_2 - \bar{x}_2)^2$
18	-19	361	29	-5	25
20	-17	389	28	-6	36
36	-1	1	26	-8	64
50	13	169	35	1	1
49	12	144	30	-4	16
36	-1	1	44	10	100
34	-3	9	46	12	144
49	12	144			
41	4	16			
333	0	1,234	238	0	386

$$\bar{x}_1 = \frac{\sum x_1}{n_1} = \frac{333}{9} = 37 \quad \text{and} \quad \bar{x}_2 = \frac{\sum x_2}{n_2} = \frac{238}{7} = 34$$

$$s = \sqrt{\frac{\sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{1234 + 386}{9 + 7 - 2}} = 10.76$$

Substituting values in the *t*-test statistic, we get

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{37 - 34}{10.76} \sqrt{\frac{9 \times 7}{9 + 7}} = \frac{3}{10.46} \times 1.984 = 0.551$$

Degrees of freedom, $df = n_1 + n_2 - 2 = 9 + 7 - 2 = 14$

Since at $\alpha = 0.05$ and $df = 14$, the calculated value $t_{\text{cal}} (= 0.551)$ is less than its critical value $t_{\alpha/2} = 2.14$, the null hypothesis H_0 is accepted. Hence we conclude that the mean marks obtained by the students of two groups do not differ significantly.

Example 8.24: The mean life of a sample of 10 electric light bulbs was found to be 1456 hours with standard deviation of 423 hours. A second sample of 17 bulbs chosen from a different batch showed a mean life of 1280 hours with standard deviation of 398 hours. Is there a significant difference between the means of the two batches.

[Delhi Univ, MCom, 1997]

Solution: Let us take the null hypothesis that there is no significant difference between the mean life of electric bulbs of two batches, that is,

$$H_0 : \mu_1 = \mu_2 \quad \text{and} \quad H_1 : \mu_1 \neq \mu_2 \quad (\text{Two-tailed test})$$

Given, $n_1 = 10$, $\bar{x}_1 = 1456$, $s_1 = 423$; $n_2 = 17$, $\bar{x}_2 = 1280$, $s_2 = 398$ and $\alpha = 0.05$. Thus,

$$\begin{aligned}\text{Pooled standard deviation, } s &= \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{9(423)^2 + 16(398)^2}{10 + 17 - 2}} \\ &= \sqrt{\frac{16,10,361 + 25,34,464}{25}} = \sqrt{1,65,793} = 407.18\end{aligned}$$

Applying the t -test, we have

$$\begin{aligned}t &= \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{1456 - 1280}{407.18} \sqrt{\frac{10 \times 17}{10 + 17}} \\ &= \frac{176}{407.18} \times 2.51 = 1.085\end{aligned}$$

Since the calculated value $t_{\text{cal}} = 1.085$ is less than its critical value $t_{\alpha/2} = 2.06$ at $df = 25$ and $\alpha = 0.05$ level of significance, the null hypothesis is accepted. Hence we conclude that the mean life of electric bulbs of two batches does not differ significantly.

Example 8.25: The manager of a courier service believes that packets delivered at the end of the month are heavier than those delivered early in the month. As an experiment, he weighed a random sample of 20 packets at the beginning of the month. He found that the mean weight was 5.25 kgs with a standard deviation of 1.20 kgs. Ten packets randomly selected at the end of the month had a mean weight of 4.96 kgs and a standard deviation of 1.15 kgs. At the 0.05 significance level, can it be concluded that the packets delivered at the end of the month weigh more?

Solution: Let us take the null hypothesis that the mean weight of packets delivered at the end of the month is more than the mean weight of packets delivered at the beginning of the month, that is

$$H_0 : \mu_E \geq \mu_B \quad \text{and} \quad H_1 : \mu_E < \mu_B$$

Given $n_1 = 20$, $\bar{x}_1 = 5.25$, $s_1 = 1.20$ and $n_2 = 10$, $\bar{x}_2 = 4.96$, $s_2 = 1.15$. Thus

$$\begin{aligned}\text{Pooled standard deviation, } s &= \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{19 \times (5.25)^2 + 9(4.96)^2}{20 + 10 - 2}} \\ &= \sqrt{\frac{19 \times 27.56 + 9 \times 24.60}{28}} = \sqrt{26.60} = 5.16\end{aligned}$$

Applying the t -test, we have

$$\begin{aligned}t &= \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{5.25 - 4.96}{5.16} \sqrt{\frac{20 \times 10}{20 + 10}} \\ &= \frac{0.29}{5.16} \sqrt{\frac{200}{30}} = 0.056 \times 2.58 = 0.145\end{aligned}$$

Since at $\alpha = 0.01$ and $df = 28$, the calculated value t_{cal} ($= 0.145$) is less than its critical value $z_{\alpha} = 1.701$, the null hypothesis is accepted. Hence, packets delivered at the end of the month weigh more on an average.

8.10.4 Hypothesis Testing for Difference of Two Population Means (Dependent Samples)

When two samples of the same size are paired so that each observation in one sample is associated with any particular observation in the second sample, the sampling procedure to collect the data and then test the hypothesis is called *matched samples*. In such a case the

'difference' between each pair of data is first calculated and then these differences are treated as a single set of data in order to consider whether there has been any significant change or whether the differences could have occurred by chance.

The matched sampling plan often leads to a smaller sampling error than the independent sampling plan because in matched samples variation as a source of sampling error is eliminated.

Let μ_d be the mean of the difference values for the population. Then this mean value μ_d is compared to zero or some hypothesized value using the *t*-test for a single sample. The *t*-test statistic is used because the standard deviation of the population of differences is unknown, and thus the statistical inference about μ_d based on the average of the sample differences \bar{d} would involve the *t*-distribution rather than the standard normal distribution. The *t*-test, also called *paired t-test*, becomes

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

where n = number of paired observations

$df = n - 1$, degrees of freedom

\bar{d} = mean of the difference between paired (or related) observations

n = number of pairs of differences

s_d = sample standard deviation of the distribution of the difference between the paired (or related) observations

$$= \sqrt{\frac{\sum (d - \bar{d})^2}{n-1}} = \sqrt{\frac{\sum d^2 - (\sum d)^2}{n(n-1)}}$$

The null and alternative hypotheses are stated as:

$$H_0 : \mu_d = 0 \text{ or } c \text{ (Any hypothesized value)}$$

$$H_1 : \mu_d > 0 \text{ or } (\mu_d < 0) \text{ (One-tailed Test)}$$

$$\mu_d \neq 0 \text{ (Two-tailed Test)}$$

Decision rule: If the calculated value t_{cal} is less than its critical value, t_d at a specified level of significance and known degrees of freedom, then null hypothesis H_0 is accepted. Otherwise H_0 is rejected.

Confidence interval: The confidence interval estimate of the difference between two population means is given by

$$\bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}}$$

where $t_{\alpha/2}$ = critical value of *t*-test statistic at $n - 1$ degrees of freedom and α level of significance.

If the claimed value of null hypothesis H_0 lies within the confidence interval, then H_0 is accepted, otherwise rejected.

Example 8.26: The HRD manager wishes to see if there has been any change in the ability of trainees after a specific training programme. The trainees take an aptitude test before the start of the programme and an equivalent one after they have completed it. The scores recorded are given below. Has any change taken place at 5 per cent significance level?

Trainee	:	A	B	C	D	E	F	G	H	I
Score before training	:	75	70	46	68	68	43	55	68	77
Score after training	:	70	77	57	60	79	64	55	77	76

Solution: Let us take the null hypothesis that there is no change that has taken place after the training, that is,

$$H_0 : \mu_d = 0 \text{ and } H_1 : \mu_d \neq 0 \text{ (Two-tailed test)}$$

The ‘changes’ are computed as shown in Table 8.6 and then a *t*-test is carried out on these differences as shown below.

Table 8.6: Calculations of ‘Changes’

Trainee	Before Training	After Training	Difference in Scores, d	d^2
A	75	70	5	25
B	70	77	7	49
C	46	57	-11	121
D	68	60	8	64
E	68	79	-11	121
F	43	64	-21	441
G	55	55	0	0
H	68	77	-9	81
I	77	76	-1	1
			-45	903

$$\bar{d} = \frac{\Sigma d}{n} = \frac{-45}{9} = -5 \text{ and}$$

$$s_d = \sqrt{\frac{\Sigma d^2}{n-1} - \frac{(\Sigma d)^2}{n(n-1)}} = \sqrt{\frac{903}{8} - \frac{(-45)^2}{9 \times 8}} = \sqrt{112.87 - 28.13} = 9.21$$

Applying the *t*-test statistic, we have

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} = \frac{-5 - 0}{9.21/\sqrt{9}} = -\frac{5}{3.07} = -1.63$$

Since the calculated value $t_{\text{cal}} = -1.63$ is more than its critical value, $t_{\alpha/2} = -2.31$, at $df = 8$ and $\alpha/2 = 0.025$ the null hypothesis is accepted. Hence, we conclude that there is no change in the ability of trainees after the training.

Example 8.27: 12 students were given intensive coaching and 5 tests were conducted in a month. The scores of tests 1 and 5 are given below.

No. of students : 1 2 3 4 5 6 7 8 9 10 11 12

Marks in 1st test : 50 42 51 26 35 42 60 41 70 55 62 38

Marks in 5th test : 62 40 61 35 30 52 68 51 84 63 72 50

Do the data indicate any improvement in the scores obtained in tests 1 and 5

[Punjab Univ., MCom, 1999]

Solution: Let us take the hypothesis that there is no improvement in the scores obtained in the first and fifth tests, that is,

$$H_0: \mu_d = 0 \text{ and } H_1: \mu_d \neq 0 \text{ (Two-tailed test)}$$

The ‘changes’ are calculated as shown in Table 8.7 and then *t*-test is carried out on these differences as shown below:

Table 8.7: Calculations of 'Changes'

No. of Students	Marks in 1st Test	Marks In 5th Test	Difference in Marks <i>d</i>	<i>d</i> ²
1	50	62	12	144
2	42	40	-2	4
3	51	61	10	100
4	26	35	9	81
5	35	30	-5	25
6	42	52	10	100
7	60	68	8	64
8	41	51	10	100
9	70	84	14	196
10	55	63	8	64
11	62	72	10	100
12	38	50	12	144
			96	1122

$$\bar{d} = \frac{\sum d}{n} = \frac{96}{12} = 8 \quad \text{and} \quad s_d = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}} = \sqrt{\frac{1122}{11} - \frac{(96)^2}{12 \times 11}} = 5.673$$

Applying the *t*-test statistic, we have

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \frac{8}{5.673 / \sqrt{12}} = 4.885$$

Since the calculated value $t_{\text{cal}} = 4.885$ is more than its critical value, $t_{\alpha/2} = 2.20$ at $df = 11$ and $\alpha/2 = 0.025$, the null hypothesis is rejected. Hence we conclude that there is an improvement in the scores obtained in two tests.

Example 8.28: To test the desirability of a certain modification in typist's desks, 9 typists were given two tests of almost same nature, one on the desk in use and the other on the new type. The following difference in the number of words typed per minute were recorded:

Typists : A B C D E F G H I	
Increase in number of words : 2 4 0 3 -1 4 -3 2 5	

Do the data indicate that the modification in desk increases typing speed?

Solution: Let us take the hypothesis that there is no change in typing speed with the modification in the typing desk, that is,

$$H_0 : \mu_d = 0 \quad \text{and} \quad H_1 : \mu_d > 0 \quad (\text{One-tailed test})$$

The calculations for 'changes' applying to the *t*-test are shown in Table 8.8.

Table 8.8: Calculations of 'Changes'

Typist	Increase in Number of Words <i>d</i>	<i>d</i> ²
A	2	4
B	4	16
C	0	0
D	3	9
E	-1	1
F	4	16
G	-3	9
H	2	4
I	5	25
	16	84

$$\bar{d} = \frac{\Sigma d}{n} = \frac{16}{9} = 1.778 \quad \text{and} \quad s_d = \sqrt{\frac{\Sigma d^2}{n-1} - \frac{(\Sigma d)^2}{n(n-1)}} = \sqrt{\frac{84}{8} - \frac{(16)^2}{9 \times 8}} = 2.635$$

Applying the *t*-test statistic, we have

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \frac{1.778}{2.625 / \sqrt{9}} = 2.025$$

Since the calculated value $t_{\text{cal}} = 2.025$ is less than its critical value $t_\alpha = 2.306$ at $df = 8$ and $\alpha = 0.05$ significance level, the null hypothesis is accepted. Hence we conclude that there is no change in typing speed with the modification in the desk.

Self-Practice Problems 8C

- 8.23** Ten oil tins are taken at random from an automatic filling machine. The mean weight of the tins is 15.8 kg and the standard deviation is 0.50 kg. Does the sample mean differ significantly from the intended weight of 16 kg? [Delhi Univ., MBA, 1998]

- 8.24** Nine items of a sample had the following values: 45, 47, 50, 52, 48, 47, 49, 53, and 50. The mean is 49 and the sum of the square of the deviation from mean is 52. Can this sample be regarded as taken from the population having 47 as mean? Also obtain 95 per cent and 99 per cent confidence limits of the population mean. [Delhi Univ., MBA, 1996]

- 8.25** The electric bulbs of 10 random samples from a large consignment gave the following data:

Item	Life in '000 hours
1	4.2
2	4.6
3	3.9
4	4.1
5	5.2
6	3.8
7	3.9
8	4.3
9	4.4
10	5.6

Can we accept the hypothesis that the average life time of the bulbs is 4000 hours. [Madras Univ., MCom, 1998]

- 8.26** A random sample of size 16 has 53 as mean. The sum of the squares of the deviations taken from mean is 135. Can this sample be regarded as taken from the population having 56 as mean? Obtain 95 per cent and 99 per cent confidence limits of the mean of the population. [Madras Univ., MCom, 1998]

- 8.27** A drug manufacturer has installed a machine which automatically fills 5 gm of drug in each phial. A random sample of fills was taken and it was found to contain 5.02 gm on an average in a phial. The standard deviation of the sample was 0.002 gms. Test at 5% level of significance if the adjustment in the machine is in order. [Delhi Univ., MBA, 1999]

- 8.28** Two salesmen *A* and *B* are working in a certain district. From a sample survey conducted by the Head Office, the following results were obtained. State whether there is any significant difference in the average sales between the two salesmen.

	<i>Salesman</i>	
	<i>A</i>	<i>B</i>
No. of samples	:	20 18
Average sales (Rs in thousand)	:	170 205
Standard deviation (Rs in thousand)	:	20 25

[Delhi Univ., MBA, 1994; Kumaon Univ., MBA, 2000]

- 8.29** The means of two random samples of sizes 9 and 7 are 196.42 and 198.82 respectively. The sum of the squares of the deviations from the mean are 26.94 and 18.73 respectively. Can the sample be considered to have been drawn from the same normal population?

[Delhi Univ., MCom, 1996]

- 8.30** Strength tests carried out on samples of two yarns spun to the same count gave the following results:

	<i>Sample size</i>	<i>Sample mean</i>	<i>Sample variance</i>
Yarn A	4	52	42
Yarn B	9	42	56

The strength is expressed in kg. Is the difference in mean strengths significant of the real difference in the mean strengths of the sources from which the samples are drawn? [Delhi Univ., MBA, 2000]

- 8.31** A random sample of 12 families in one city showed an average monthly food expenditure of Rs 1380 with a standard deviation of Rs 100 and a random sample of 15 families in another city showed an average monthly food expenditure of Rs 1320 with a standard deviation of Rs 120. Test whether the difference between the two means is significant at $\alpha = 0.01$ level of significance of $\alpha = 0.01$.

[AIMA Diploma in Mgt., 1987; Delhi Univ., MBA, 1991]

- 8.32** You are given the following data about the life of two brands of bulbs:

	<i>Mean life</i>	<i>Standard deviation</i>	<i>Sample size</i>
Brand A	2000 hrs	250 hrs	12
Brand B	2230 hrs	300 hrs	15

Do you think there is a significant difference in the quality of the two brands of bulbs?

[Delhi Univ., MBA, 1996]

- 8.33** Eight students were given a test in statistics, and after one month's coaching, they were given another test of the similar nature. The following table gives the increase in their marks in the second test over the first:

<i>Roll No.</i>	<i>Increase in marks</i>
1	2
2	-2
3	6
4	-8
5	12
6	5
7	-7
8	2

Do the marks indicate that the students have gained from the coaching?

- 8.34** An IQ test was administered to 5 persons before and after they were trained. The results are given below:

<i>Candidate</i>	<i>I</i>	<i>II</i>	<i>III</i>	<i>IV</i>	<i>V</i>
IQ before training :	110	120	123	132	125
IQ after training :	120	118	.25	136	121

Test whether there is any change in IQ level after the training programme. [Delhi Univ., MCom, 1998]

- 8.35** Eleven sales executive trainees are assigned selling jobs right after their recruitment. After a fortnight they are withdrawn from their field duties and given a month's training for executive sales. Sales executed by them in thousands of rupees before and after the training, in the same period are listed below:

<i>Sales Before Training</i>	<i>Sales After Training</i>
23	24
20	19
19	21
21	18
18	20
20	22
18	20
17	20
23	23
16	20
19	27

Do these data indicate that the training has contributed to their performance? [Delhi Univ., MCom, 1999]

Hints and Answers

- 8.23** Let $H_0 : \mu = 16$ and $H_1 : \mu \neq 16$ (Two-tailed test)
Given $n = 10$, $\bar{x} = 15.8$, $s = 0.50$. Using t -test

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{15.8 - 16}{0.50/\sqrt{10}} = -1.25$$

Since $t_{\text{cal}} = -1.25 >$ critical value $t_{\alpha/2} = -2.262$, at $df = 9$ and $\alpha/2 = 0.025$, the null hypothesis is accepted.

- 8.24** Let $H_0 : \mu = 27$ and $H_1 : \mu \neq 47$ (Two-tailed test)
Given $\bar{x} = 49$, $\Sigma(x - \bar{x})^2 = 52$, $n = 9$, and

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{(n-1)}} = \sqrt{\frac{52}{8}} = 2.55.$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{49 - 47}{2.55/\sqrt{9}} = 2.35$$

Since $t_{\text{cal}} = 2.35 >$ critical value $t_{\alpha/2} = 2.31$ at $\alpha/2 = 0.025$, $df = 8$, the null hypothesis is rejected.

- 8.25** Let $H_0 : \mu = 4,000$ and $H_1 : \mu \neq 4000$ (Two-tailed test)

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{(n-1)}} = \sqrt{\frac{3.12}{9}} = 0.589 \text{ and}$$

$$\bar{x} = \sum x/n = 44/10 = 4.4 \text{ (in Rs 000's).}$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{4.4 - 4}{0.589/\sqrt{10}} = \frac{0.4}{0.186} = 2.150$$

Since $t_{\text{cal}} = 2.150 <$ critical value $t_{\alpha/2} = 2.62$ at $\alpha/2 = 0.025$ and $df = n-1 = 9$, the null hypothesis is accepted.

- 8.26** Let $H_0 : \mu = 56$ and $H_1 : \mu \neq 46$ (Two-tailed test)
Given: $n = 16$, $\bar{x} = 53$ and $\Sigma(x - \bar{x})^2 = 135$. Thus

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{(n-1)}} = \sqrt{\frac{135}{15}} = 3$$

$$\text{Applying } t\text{-test, } t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{53 - 56}{3/\sqrt{16}} = -4$$

Since $t_{\text{cal}} = -4 <$ critical value $t_{\alpha/2} = -2.13$ at $\alpha/2 = 0.025$, $df = 15$, the null hypothesis is rejected.

- 8.27** Let $H_0 : \mu = 5$ and $H_1 : \mu \neq 5$ (Two-tailed test)
Given $n = 10$, $\bar{x} = 5.02$ and $s = 0.002$.

$$\text{Applying } t\text{-test, } t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{5.02 - 5}{0.002/\sqrt{10}} = 33.33$$

Since $t_{\text{cal}} = 33.33 >$ critical value $t_{\alpha/2} = 1.833$ at $\alpha/2 = 0.025$ and $df = 9$, the null hypothesis is rejected.

- 8.28** Let $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$ (Two-tailed test)
Given $n_1 = 20$, $s_1 = 20$, $\bar{x}_1 = 170$; $n_2 = 18$, $s_2 = 25$, $\bar{x}_2 = 205$, Applying t -test,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s/\sqrt{n}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$= \frac{170 - 205}{22.5} \sqrt{\frac{20 \times 18}{20 + 18}} = \frac{-35}{22.5} \sqrt{\frac{360}{38}} = -4.8$$

$$\begin{aligned}s &= \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \\&= \sqrt{\frac{19(20)^2 + 17(25)^2}{20+18-2}} = \sqrt{\frac{18,225}{36}} = 22.5\end{aligned}$$

Since $t_{\text{cal}} = -4.8 <$ critical value $t_{\alpha/2} = -1.9$ at $\alpha/2 = 0.025$ and $df = n_1 + n_2 - 2 = 36$, the null hypothesis is rejected.

- 8.29** Let $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$ (Two-tailed test)

Given $n_1 = 9$, $\bar{x}_1 = 196.42$, $\Sigma(x_1 - \bar{x}_1)^2 = 26.94$ and $n_2 = 7$, $\bar{x}_2 = 198.82$ and $\Sigma(x_2 - \bar{x}_2)^2 = 18.73$

$$\begin{aligned}t &= \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{196.42 - 198.82}{1.81} \sqrt{\frac{9 \times 7}{9 + 7}} \\&= -\frac{2.40}{1.81} \sqrt{\frac{63}{16}} = -2.63 \\s &= \sqrt{\frac{\Sigma (x_1 - \bar{x}_1)^2 + (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}} \\&= \sqrt{\frac{26.94 + 18.73}{9 + 7 - 2}} = 1.81\end{aligned}$$

Since $t_{\text{cal}} = -2.63 <$ critical value $t_{\alpha/2} = -2.145$ at $\alpha/2 = 0.025$ and $df = 14$, the null hypothesis is rejected.

- 8.30** Let $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$ (Two-tailed test)

Given $n_1 = 4$, $s_1^2 = 42$, $\bar{x}_1 = 52$, and $n_2 = 9$, $s_2^2 = 56$, $\bar{x}_2 = 42$.

$$\begin{aligned}t &= \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{52 - 42}{7.224} \sqrt{\frac{4 \times 9}{4 + 9}} \\&= \frac{10}{7.224} \sqrt{\frac{36}{13}} = 2.303 \\s &= \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \\&= \sqrt{\frac{3 \times 42 + 8 \times 56}{4 + 9 - 2}} = \sqrt{\frac{574}{11}} = 7.224\end{aligned}$$

Since $t_{\text{cal}} = 2.303 >$ critical value $t_{\alpha/2} = 2.20$ at $\alpha/2 = 0.025$ and $df = n_1 + n_2 - 2 = 11$, the null hypothesis is rejected.

- 8.31** Let $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$ (Two-tailed test)

Given $n_1 = 12$, $s_1 = 100$, $\bar{x}_1 = 1380$ and $n_2 = 15$, $s_2 = 120$, $\bar{x}_2 = 1320$. Applying t -test,

$$\begin{aligned}s &= \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \\&= \sqrt{\frac{11(100)^2 + 14(120)^2}{12+15-2}} = 111.64\end{aligned}$$

$$\begin{aligned}t &= \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{1380 - 1320}{111.64} \sqrt{\frac{12 \times 15}{12 + 15}} \\&= \frac{60}{111.64} \sqrt{\frac{180}{27}} = 1.39\end{aligned}$$

Since $t_{\text{cal}} = 1.39 <$ critical value $t_{\alpha/2} = 2.485$ at $\alpha/2 = 0.025$ and $df = n_1 + n_2 - 2 = 25$, the null hypothesis is accepted.

- 8.32** Let $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$ (Two-tailed test)

Given $n_1 = 12$, $\bar{x}_1 = 2000$, $s_1 = 250$ and $n_2 = 15$, $\bar{x}_2 = 2230$, $s_2 = 300$. Applying t -test,

$$\begin{aligned}s &= \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \\&= \sqrt{\frac{11(250)^2 + 14(300)^2}{12+15-2}} = 279.11 \\t &= \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \\&= \frac{2000 - 2230}{279.11} \sqrt{\frac{12 \times 15}{12 + 15}} \\&= -\frac{230}{279.11} \sqrt{\frac{180}{27}} = -2.126\end{aligned}$$

Since $t_{\text{cal}} = -2.126 <$ critical value $t_{\alpha/2} = -1.708$ at $\alpha/2 = 0.025$ and $df = n_1 + n_2 - 2 = 25$, the null hypothesis is rejected.

- 8.33** Let H_0 : Students have not gained from coaching

Roll No	Increase in marks, d	d^2
1	4	16
2	-2	4
3	6	36
4	-8	64
5	12	144
6	5	25
7	-7	49
8	2	4
	12	342

$$\bar{d} = \frac{\sum d}{n} = \frac{12}{8} = 1.5;$$

$$s = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}} = \sqrt{\frac{342}{7} - \frac{12^2}{8 \times 7}} = 6.8$$

$$t = \frac{\bar{d}}{s/\sqrt{n}} = \frac{1.5}{6.8/\sqrt{8}} = 0.624$$

Since $t_{\text{cal}} = 0.622 <$ critical value $t_{\alpha/2} = 1.895$ at $\alpha/2 = 0.025$ and $df = 7$, null hypothesis is accepted.

8.34 Let H_0 : No change in IQ level after training programme

<i>IQ Level</i>		<i>d</i>	<i>d</i> ²
<i>Before</i>	<i>After</i>		
110	120	10	100
120	118	-2	4
123	125	2	4
132	136	4	16
125	121	-4	16
		10	140

$$\bar{d} = \frac{\sum d}{n} = \frac{10}{5} = 2$$

$$s = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}} = \sqrt{\frac{140}{4} - \frac{(10)^2}{5 \times 4}} = 5.477$$

$$t = \frac{\bar{d}}{s/\sqrt{n}} = \frac{2}{5.477/\sqrt{5}} = 0.817$$

Since $t_{\text{cal}} = 0.814 <$ critical value $t_{\alpha/2} = 4.6$ at $\alpha/2 = 0.025$ and $df = 4$, the null hypothesis is accepted.

8.35 Let H_0 : Training did not improve the performance of the sales executives

<i>Difference in Sales, d</i>	<i>d</i> ²
1	1
-1	1
2	4
-3	9
2	4
2	4
2	4
3	9
0	0
4	16
8	64
= 20	= 116

$$\bar{d} = \frac{\sum d}{n} = \frac{20}{11} = 1.82$$

$$s = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}} = \sqrt{\frac{116}{10} - \frac{(20)^2}{11 \times 10}} = 2.82$$

$$t = \frac{\bar{d}}{s/\sqrt{n}} = \frac{1.82}{2.82/\sqrt{11}} = 2.14$$

Since $t_{\text{cal}} = 2.14 <$ critical value $t_{\alpha/2} = 2.23$ at $\alpha/2 = 0.025$ and $df = 10$, the null hypothesis is accepted.

8.11 HYPOTHESIS TESTING BASED ON F-DISTRIBUTION

In several statistical applications we might require to compare population variances. For instance, (i) variances in product quality resulting from two different production processes; (ii) variances in temperatures for two heating devices; (iii) variances in assembly times for two assembly methods, (iv) variance in the rate of return on investment of two types of stocks and so on, are few areas where comparison of variances is needed.

When independent random samples of size n_1 and n_2 are drawn from two normal populations, the ratio

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

follow F-distribution with $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$ degrees of freedom, where s_1^2 and s_2^2 are two sample variances and are given by

$$s_1^2 = \frac{\sum (x_1 - \bar{x}_1)^2}{n_1 - 1} \text{ and } s_2^2 = \frac{\sum (x_2 - \bar{x}_2)^2}{n_2 - 1}$$

If two normal populations have equal variances, i.e. $\sigma_1^2 = \sigma_2^2$, then the ratio

$$F = \frac{s_1^2}{s_2^2}; s_1 > s_2$$

F-test: A hypothesis test for comparing the variance of two independent populations with the help of variances of two small samples.

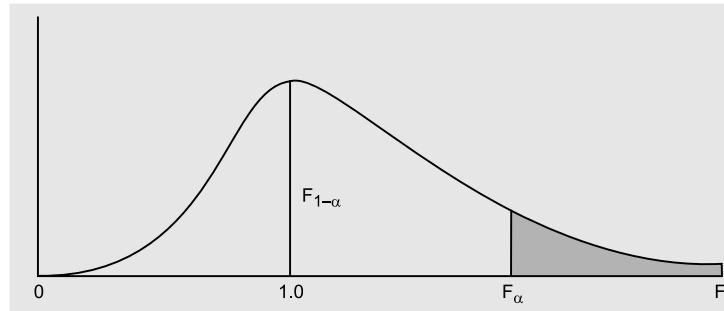
has a probability distribution in repeated sampling that is known as F-distribution with $n_1 - 1$ degrees of freedom for numerator and $n_2 - 1$ degrees of freedom for denominator. For computational purposes, a larger sample variance is placed in the numerator so that ratio is always equal to or more than one.

Assumptions: Few assumptions for the ratio s_1^2/s_2^2 to have an F-distribution are as follows:

- (i) Independent random samples are drawn from each of two normal populations
- (ii) The variability of the measurements in the two populations is same and can be measured by a common variance σ^2 , i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$

The F-distribution, also called *variance ratio distribution*, is not symmetric and the F values can never be negative. The shape of any F-distribution depends on the degrees of freedom of the numerator and denominator. A typical graph of a F-distribution is shown in Fig. 8.10 for equal degrees of freedom for both numerator and denominator.

Figure 8.10
F-distribution for n Degrees
of Freedom



8.11.1 Properties of F-distribution

1. The total area or probability under the curve is unity. The value of F-test statistic denoted by F_α at a particular level of significance α , provides an area or probability of α to the right of the stated F_α value.
2. The F-distribution is positively skewed with a range 0 to ∞ and its degree of skewness decreases with the increase in degrees of freedom v_1 for numerator and v_2 for denominator. For $v_2 \geq 30$, F-distribution is approximately normal.
3. The sample variances s_1^2 and s_2^2 are the unbiased estimates of population variance. Since $s_1 > s_2$, the range of F-distribution curve is from 1 to ∞ .
4. If the ratio s_1^2/s_2^2 is nearly equal to 1, then it indicates little evidence that σ_1^2 and σ_2^2 are unequal. On the other hand, a very large or very small value for s_1^2/s_2^2 provides evidence of difference in the population variances.
5. The F-distribution discovered by Sir Ronald Fisher is merely a transformation of the original Fisher's z-distribution and is written as

$$\begin{aligned} z &= \frac{1}{2} \log_e F = \frac{1}{2} \log_e \frac{s_1^2}{s_2^2} = \frac{1}{2} \log_e \left(\frac{s_1}{s_2} \right)^2 \\ &= \log_e \left(\frac{s_1}{s_2} \right) = \log_{10} \left(\frac{s_1}{s_2} \right) \log_e 10 = 2.3026 \log_{10} \left(\frac{s_1}{s_2} \right) \end{aligned}$$

The probability density function of F-distribution is given by

$$f(F) \text{ or } y = y_0 \frac{e^{v_1 z}}{(v_1 e^{2z} + v_2)^{(v_1 + v_2)/2}}, \quad -\infty < z < \infty$$

where y_0 is a constant depending on the values of degrees of freedom, $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$, such that the area under the curve is unity.

6. The mean and variances of the F-distribution are

$$\text{Mean } \mu = \frac{v_2}{v_2 - 2}, \quad v_2 > 2$$

$$\text{and} \quad \text{Variance } \sigma^2 = \frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - v_1)^2(v_2 - 4)}, \quad v_2 > 4$$

This implies that F-distribution has no mean for $v_2 \leq 2$ and no variance for $v_2 \leq 4$.

7. The reciprocal property

$$F_{1-\alpha(v_2, v_1)} = \frac{1}{F_{\alpha(v_2, v_1)}}$$

of F-distribution helps to identify corresponding lower (left) tail F values from the given upper (right) tail F values. For example, if $\alpha = 0.05$, then for $v_1 = 24$ and $v_2 = 15$, then $F_{0.95(15, 24)} = 2.11$. Thus $F_{0.95(24, 15)} = 1/2.11 = 0.47$.

8. The variance of F with 1 and n degrees of freedom is distributed same as t -distribution with n degrees of freedom.

8.11.2 Comparing Two Population Variances

How large or small must the ratio s_1^2/s_2^2 be for sufficient evidence to exist to the null hypothesis is stated below:

Null hypothesis	Alternative hypothesis
$H_0: \sigma_1^2 = \sigma_2^2$	$H_1: \sigma_1^2 > \sigma_2^2$ or $\sigma_1^2 < \sigma_2^2$ (One-tailed Test)
$H_0: \sigma_1^2 = \sigma_2^2$	$H_1: \sigma_1^2 \neq \sigma_2^2$ (Two-tailed Test)

To conduct the test, random samples of size n_1 and n_2 are drawn from population 1 and 2 respectively. The statistical test of the null hypothesis H_0 , uses the test statistic $F = s_1^2/s_2^2$, where s_1^2 and s_2^2 are the respective sample variances.

Decision rules: The criteria of acceptance or rejection of null hypothesis H_0 are as under:

1. Accept H_0 if the calculated value of F-test statistic is less than its critical value $F_{\alpha(v_1, v_2)}$, i.e. $F_{\text{cal}} < F_{\alpha}$ for one-tailed test.

The critical value of F_{α} is based on degrees of freedom of numerator $df_1 = n_1 - 1$ and degrees of freedom of denominator $df_2 = n_2 - 1$. These values can be obtained from F-Tables (See appendix).

As mentioned earlier, the population with larger variance is considered as population 1 to ensure that a rejection of H_0 can occur only in the right (upper) tail of the F-distribution curve. Even though half of the rejection region (the area $\alpha/2$ to its left) will be in the lower tail of the distribution. It is never used because using the population with larger sample variance as population 1 always places the ratio s_1^2/s_2^2 in the right-tail direction.

2. $H_0: \sigma_1^2 = \sigma_2^2$ and $H_1: \sigma_1^2 > \sigma_2^2$ (One-tailed test)

The null hypothesis is setup so that the rejection region is always in the upper tail of the distribution. This helps us in considering the population with larger variance in the alternative hypothesis.

Confidence Interval: An interval estimate of all possible values for a ratio σ_1^2/σ_2^2 of population variances, is given by

$$\frac{s_1^2/s_2^2}{F_{(1-\alpha)}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2/s_2^2}{F_{\alpha}}$$

where F values are based on a F-distribution with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom and $(1 - \alpha)$ confidence coefficient.

Example 8.29: A research was conducted to understand whether women have a greater variation in attitude on political issues than men. Two independent samples of 31 men and 41 women were used for the study. The sample variances so calculated were 120 for women and 80 for men. Test whether the difference in attitude toward political issues is significant at 5 per cent level of significance.

Solution: Let us take the hypothesis that the difference is not significant, that is,

$$H_0: \sigma_w^2 = \sigma_m^2 \quad \text{and} \quad H_1: \sigma_w^2 > \sigma_m^2 \quad (\text{One-tailed test})$$

$$\text{The F-test statistic is given by } F = \frac{s_1^2}{s_2^2} = \frac{120}{80} = 1.50$$

Since variance for women is in the numerator, the F-distribution with $df_1 = 41 - 1 = 40$ in the numerator and $df_2 = 31 - 1 = 30$ in the denominator will be used to conduct the one-tailed test.

The critical (table) value of $F_{\alpha=0.05} = 1.79$ at $df_1 = 40$ and $df_2 = 30$. The calculated value of $F = 1.50$ is less than its critical value $F = 1.79$, the null hypothesis is accepted. Hence, the results of the research do not support the belief that women have a greater variation in attitude on political issues than men.

Example 8.30: The following figures relate to the number of units of an item produced per shift by two workers A and B for a number of days

A:	19	22	24	27	24	18	20	19	25
B:	26	37	40	35	30	30	40	26	30

Can it be inferred that worker A is more stable compared to worker B? Answer using the F-test at 5 per cent level of significance.

Solution: Let us take the hypothesis that the two workers are equally stable, that is,

$$H_0: \sigma_A^2 = \sigma_B^2 \text{ and } H_1: \sigma_A^2 \neq \sigma_B^2 \text{ (One-tailed test)}$$

The calculations for population variances σ_A^2 and σ_B^2 of the number of units produced by workers A and B, respectively are shown in Table 8.9.

Table 8.9: Calculation of σ_A^2 and σ_B^2

Worker A	$x_I - \bar{x}_I$	$(x_I - 22)^2$	Worker B	$x_2 - \bar{x}_2$	$(x_2 - \bar{x}_2)^2$
	$x_I = x_I - 22$			$x_2 = x_2 - 34$	
19	-3	9	26	-8	64
22	0	0	37	3	9
24	2	4	40	6	36
27	5	25	35	1	1
24	2	4	30	-4	16
18	-4	16	30	-4	16
20	-2	4	40	6	36
19	-3	9	26	-8	64
25	3	9	30	-4	16
			35	1	1
<u>198</u>	<u>0</u>	<u>80</u>	<u>374</u>	<u>11</u>	<u>121</u>
					<u>380</u>

$$\bar{x}_1 = \frac{\sum x_1}{n_1} = \frac{198}{9} = 22;$$

$$\bar{x}_2 = \frac{\sum x_2}{n_2} = \frac{374}{11} = 34$$

$$s_A^2 = \frac{\sum (x_1 - \bar{x}_1)^2}{n_1 - 1} = \frac{80}{9-1} = 10;$$

$$s_B^2 = \frac{\sum (x_2 - \bar{x}_2)^2}{n_2 - 1} = \frac{380}{11-1} = 38$$

Applying F-test statistic, we have

$$F = \frac{s_B^2}{s_A^2} = \frac{38}{10} = 3.8 \text{ (because } s_B^2 > s_A^2\text{)}$$

The critical value $F_{0.05(10, 8)} = 3.35$ at $\alpha = 5$ per cent level of significance and degrees of freedom $df_A = 8$, $df_B = 10$. Since the calculated value of F is more than its critical value, the null hypothesis is rejected. Hence we conclude that worker A is more stable than worker B, because $s_A^2 < s_B^2$.

Self-Practice Problems 8D

- 8.36** The mean diameter of a steel pipe produced by two processes, A and B, is practically the same but the standard deviations may differ. For a sample of 22 pipes produced by A, the standard deviation is 2.9 m, while for a sample of 16 pipes produced by B, the standard deviation is 3.8 m. Test whether the pipes produced by process A have the same variability as those of process B.

- 8.37** Tests for breaking strength were carried out on two lots of 5 and 9 steel wires respectively. The variance of one lot was 230 and that of the other was 492. Is there a significant difference in their variability?

- 8.38** Two random samples drawn from normal population are:

Sample 1	Sample 2
20	27
16	33
26	42
27	35
23	32
22	34
18	38
24	28
25	41
19	43
	30
	37

Obtain estimates of the variances of the population and test whether the two populations have the same variance.

- 8.39** In a sample of 8 observations, the sum of the squared deviations of items from the mean was 94.50. In another sample of 10 observations the value was found to be 101.70. Test whether the difference is significant at

5 per cent level of significance (at 5 per cent level level of significance critical value of F for $v_1 = 3$ and $v_2 = 9$ degrees of freedom is 3.29 and for $v_1 = 8$ and $v_2 = 10$ degrees of freedom, its value is 3.07).

- 8.40** Most individuals are aware of the fact that the average annual repair costs for an automobile depends on the age of the automobile. A researcher is interested in finding out whether the variance of the annual repair costs also increases with the age of the automobile. A sample of 25 automobiles that are 4 years old showed a sample variance for annual repair cost of Rs 850 and a sample of 25 automobiles that are 2 years old showed a sample variance for annual repair costs of Rs 300. Test the hypothesis that the variance in annual repair costs is more for the older automobiles, for a 0.01 level of significance.

- 8.41** The standard deviation in the 12-month earnings per share for 10 companies in the software industry was 4.27 and the standard deviation in the 12-month earning per share for 7 companies in the telecom industry was 2.27. Conduct a test for equal variance at $\alpha = 0.05$. What is your conclusion about the variability in earning per share for two industries.

Hints and Answers

- 8.36** Let H_0 : There is no difference in the variability of diameters produced by process A and B, i.e.

$$H_0 : \sigma_A^2 = \sigma_B^2 \text{ and } H_1 : \sigma_A^2 \neq \sigma_B^2$$

Given $\sigma_A = 2.9$, $n_1 = 22$, $df_A = 21$; $\sigma_B = 3.8$, $n_2 = 16$, $df_B = 21$.

$$s_A^2 = \frac{n_1}{n_1 - 1} \sigma_A^2 = \frac{22}{22 - 1} (2.9)^2 = \frac{22}{21} (8.41) = 8.81$$

$$s_2^2 = \frac{n_2}{n_2 - 1} \sigma_B^2 = \frac{16}{16 - 1} (3.8)^2 = \frac{16}{15} (14.44) = 15.40$$

$$F = \frac{s_2^2}{s_1^2} = \frac{15.40}{8.81} = 1.75$$

Since the calculated value $F = 1.75$ is less than its critical value $F_{0.05(15, 21)} = 2.18$, the null hypothesis is accepted.

- 8.37** Let H_0 : No significant variability in the breaking strength of wires

Given $n_1 = 5$, $\sigma_1^2 = 230$, $df_1 = 4$; $n_2 = 9$, $\sigma_2^2 = 492$, $df_2 = 8$

$$F = \frac{\sigma_2^2}{\sigma_1^2} = \frac{492}{230} = 2.139$$

Since calculated value $F = 2.139$ is less than its critical value $F_{0.05(8, 4)} = 6.04$ the null hypothesis is accepted.

- 8.38** Let H_0 : Two populations have the same variance, i.e.

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ and } H_1 : \sigma_1^2 \neq \sigma_2^2.$$

$$\text{Sample 1: } \bar{x}_1 = \frac{\sum x_1}{10} = 22;$$

$$s_1^2 = \frac{\sum (x_1 - \bar{x}_1)^2}{n_1 - 1} = \frac{120}{9} = 13.33, df_1 = 9$$

$$\text{Sample 2: } \bar{x}_2 = \frac{\sum x_2}{12} = 35;$$

$$s_2^2 = \frac{\sum (x_2 - \bar{x}_2)^2}{n_2 - 1} = \frac{314}{11} = 28.54, df_2 = 11$$

$$F = \frac{s_2^2}{s_1^2} = \frac{28.54}{13.33} = 2.14$$

Since calculated value $F = 2.14$ is less than its critical value $F_{0.05(11, 9)} = 4.63$, the null hypothesis is accepted.

- 8.39** Let H_0 : The difference is not significant

Sample 1: $n_1 = 8$, $\sum (x_1 - \bar{x}_1)^2 = 94.50$, $v_1 = 7$

Sample 2: $n_1 = 10$, $\sum (x_2 - \bar{x}_2)^2 = 101.70$, $v_2 = 9$

$$\therefore s_1^2 = \frac{\sum (x_1 - \bar{x}_1)^2}{n_1 - 1} = \frac{94.50}{7} = 13.5;$$

$$s_2^2 = \frac{\sum (x_2 - \bar{x}_2)^2}{n_2 - 1} = \frac{101.70}{9} = 11.3$$

$$F = \frac{s_1^2}{s_2^2} = \frac{13.5}{11.3} = 1.195$$

Since calculated value $F = 1.195$ is less than its critical value $F_{0.05(7, 9)} = 3.29$, the null hypothesis is accepted.

- 8.40** Let H_0 : No significant difference in the variance of repair cost, $H_0 : \sigma_1^2 = \sigma_2^2$ and $H_1 : \sigma_1^2 > \sigma_2^2$

$$s_1^2 = \text{Rs } 850; s_2^2 = \text{Rs } 300$$

$$n_1 = 25, df_1 = 24; n_2 = 25, df_2 = 24$$

$$F = \frac{s_1^2}{s_2^2} = \frac{850}{300} = 2.833$$

Since the calculated value $F = 2.833$ is more than its critical value $F_{0.01(24, 24)} = 2.66$, the null hypothesis is rejected.

- 8.41** Let H_0 : No significant difference of variability in earning per share for two industries,

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ and } H_1 : \sigma_1^2 \neq \sigma_2^2$$

Software industry : $s_1^2 = (4.27)^2 = 18.23$,

$$n_1 = 10, df_1 = 9$$

Telecom industry : $s_2^2 = (2.27)^2 = 5.15$,

$$n_2 = 7, df_2 = 6$$

$$\therefore F = \frac{s_1^2}{s_2^2} = \frac{18.23}{5.15} = 3.54$$

Since the calculated value $F = 3.54$ is less than its critical value $F_{0.05(9, 6)} = 4.099$, the null hypothesis is accepted.

Formulae Used

1. Hypothesis testing for population mean with large sample ($n > 30$)

- (a) Test statistic about a population mean μ

- σ assumed known, $z = \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}}$

- σ is estimated by s , $z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

- (b) Test statistic for the difference between means of two populations

- Standard deviation of $\bar{x}_1 - \bar{x}_2$ when σ_1 and σ_2 are known

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\text{Test statistic } z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

- Standard deviation of $\bar{x}_1 - \bar{x}_2$ when $\sigma_1^2 = \sigma_2^2$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- Point estimator of $\sigma_{\bar{x}_1 - \bar{x}_2}$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Interval estimation for single population mean

$$\bar{x} \pm z_{\alpha/2} \sigma_{\bar{x}} ; \sigma \text{ is known}$$

$$\bar{x} \pm z_{\alpha/2} s_{\bar{x}} ; \sigma \text{ is unknown}$$

- Interval estimation for the difference of means of two populations

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sigma_{\bar{x}_1 - \bar{x}_2} ; \sigma_1 \text{ and } \sigma_2 \text{ are known}$$

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} s_{\bar{x}_1 - \bar{x}_2} ; \sigma_1 \text{ and } \sigma_2 \text{ are unknown}$$

2. Hypothesis testing for population proportion for large sample ($n > 30$)

- (a) Test statistic for population proportion p

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} ; \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

- (b) Test statistic for the difference between the proportions of two populations

- Standard deviation of $\bar{p}_1 - \bar{p}_2$

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- Point estimator of $\sigma_{\bar{p}_1 - \bar{p}_2}$

$$s_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$$

- Interval estimation of the difference between the proportions of two populations

$$(\bar{p}_1 - \bar{p}_2) \pm z_{\alpha/2} s_{\bar{p}_1 - \bar{p}_2}$$

where all $n_1 p_1, n_1(1-p_1), n_2 p_2$, and $n_2(1-p_2)$ are more than or equal to 5

- Test statistic for hypothesis testing about the difference between proportions of two

$$\text{populations } z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sigma_{\bar{p}_1 - \bar{p}_2}}$$

- Pooled estimator of the population proportion

$$\bar{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2}$$

- Point estimator of $\sigma_{\bar{p}_1 - \bar{p}_2}$

$$s_{\bar{p}_1 - \bar{p}_2} = \sqrt{\bar{p}(1-\bar{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

3. Hypothesis testing for population mean with small sample ($n \leq 30$)

- Test statistic when s is estimated by s

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where $s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

$$= \sqrt{\frac{\sum (x_1 - \bar{x}_1)^2 + (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

- Test statistic for difference between the means of two population proportions

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

where $s_{\bar{x}_1 - \bar{x}_2}$ is the point estimator of $\sigma_{\bar{p}_1 - \bar{p}_2}$ when $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- Interval estimation of the difference between means of two populations $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_{\bar{x}_1 - \bar{x}_2}$

4. Hypothesis testing for matched samples (small sample case)

Test statistic for matched samples

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} ; \quad s_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n-1}} ; \quad \bar{d} = \frac{\sum d}{n}$$

5. Hypothesis testing for two population variances

$$F = s_1^2 / s_2^2 ; \quad s_1^2 > s_2^2$$

Review-Self Practice Problems

- 8.42** A sample of size 25 yielded a mean equal to 33 and an estimated variance equal to 100. At the $\alpha = 0.01$ would we have reasons to doubt the claim that the population mean is not greater than 27?

[Kurukshetra Univ., MCom, 1996]

- 8.43** An educator claims that the average IQ of American college students is at most 110, and that in a study made to test this claim 150 American college students selected at random had an average IQ of 111.2 with a standard deviation of 7.2. Use a level of significance of 0.01 to test the claim of the educator.

[Delhi Univ., BA (H.Eco.), 1996]

- 8.44** Prices of shares (in Rs) of a company on different days in a month were found to be: 66, 65, 69, 70, 69, 71, 70, 63, 64, and 68. Test whether the mean price of the shares in the month is 65. [Delhi Univ., MBA, 1999]

- 8.45** 500 apples are taken at random from a large basket and 50 are found to be bad. Estimate the proportion of bad apples in the basket and assign limits within which the percentage most probably lies.

- 8.46** The election returns showed that a certain candidate received 46 per cent of the votes. Determine the probability that a poll of (a) 200 and (b) 1000 people selected at random from the voting population would have shown a majority of votes in favour of the candidate.

- 8.47** A simple random sample of size 100 has mean 15, the population variance being 25. Find an interval estimate of the population mean with a confidence level of (i) 99 per cent, and (ii) 95 per cent. If population variance is not given, what should be known to find out the required interval estimates? [CA Nov., 1988]

- 8.48** A machine produced 20 defective articles in a batch of 400. After overhauling, it produced 10 defectives in a batch of 300. Has the machine improved?

[Madras Univ., MCom, 1996]

- 8.49** You are working as a purchase manager for a company. The following information has been supplied to you by two manufacturers of electric bulbs:

Company A Company B

Mean life (in hours)	:	1300	1288
Standard deviation (in hours)	:	82	93
Sample size	:	100	100

Which brand of bulbs you will prefer to purchase if your desire is to take a risk of 5 per cent.

[Delhi Univ., MCom, 1994; MBA, 1998, 2002]

- 8.50** The intelligence test given to two groups of boys and girls gave the following information:

	Mean Score	Standard Deviation	Number
Girls	75	10	50
Boys	70	12	100

Is the difference in the mean scores of boys and girls statistically significant? [Delhi Univ., MBA, 1997]

- 8.51** Two samples of 100 electric bulbs each has a mean length of life 1500 and 1550 hours and standard deviation of 50 and 60 hours. Can it be concluded that two brands differ significantly at 1 per cent level of significance in equality? [Kurukshetra Univ., MBA, 1999]

- 8.52** Two types of drugs were used on 5 and 7 patients for reducing their weight. Drug A was imported and drug B indigenous. The decrease in the weights after using the drugs for six months was as follows:

Drug A :	10	12	13	11	14	10	9
Drug B :	8	9	12	14	15		

Is there a significant difference in the efficacy of the two drugs? If not, which drug should you buy?

[Osmania Univ., MBA, 1996]

- 8.53** Samples of two different types of bulbs were tested for length of life, and the following data were obtained:

	Type I	Type II
Sample size	:	8
Sample mean	:	1234 hrs
Sample S.D.	:	36 hrs
		40 hrs

Is the difference in mean life of two types of bulbs significant? [Delhi Univ., MBA, 1996]

- 8.54** A college conducting both day and evening classes intends them to be identical. A sample of 100 day students yields examination results as: $\bar{x}_1 = 72.4$, $\sigma_1 = 14.8$. Similarly, a sample of 200 evening students yields examination results as: $\bar{x}_2 = 73.9$, $\sigma_2 = 17.9$. Are

the day and evening classes statistically equal at $\alpha = 0.01$ level of significance? [Sukhadia Univ., M.Com, 1989]

- 8.55** A strength test carried out on samples of two yarns spun to the same count gave the following results:

	Sample Size	Sample Mean	Sample Variance
Yarn A	4	52	42
Yarn B	9	42	56

The strengths are expressed in pounds. Is the difference in mean strengths significant of the sources from which the samples are drawn? [Delhi Univ., MBA, 1998]

- 8.56** An investigation of the relative merits of two kinds to flashlight batteries showed that a random sample of 100 batteries of brand X lasted on the average 36.5 hours with a standard deviation of 1.8 hours, while a random sample of 80 batteries of brand Y lasted on the average 36.8 hours with a standard deviation of 1.5 hours. Use a level of significance of 0.05 to test whether the observed difference between the average lifetimes is significant. [Delhi Univ., BA (H. Econ.), 1996]

- 8.57** A company is interested in finding out if there is any difference in the average salary received by the managers of two divisions. Accordingly, samples of 12 managers of the first division and 10 managers of the second division were selected at random. The results are given below:

	First Division	Second Division
Sample size	12	10
Average monthly salary	12,500	11,200
Standard deviation	320	480

Apply the *t*-test to find out whether there is a significant difference in the average salary.

[Kumaon Univ., MBA, 1999]

- 8.58** A random sample of 100 mill workers in Kanpur showed their mean wage to be Rs 3500 with a standard deviation of Rs 280. Another random sample of 150 mill workers in Mumbai showed the mean wage to be Rs 3900 with a standard deviation of Rs 400. Do the mean wages of workers in Mumbai and Kanpur differ significantly? Use $\alpha = 0.05$ level of significance.

[Delhi Univ., MCom, 1999]

- 8.59** The sales data of an item in six shops before and after a special promotional campaign are as under:

Shops	:	A	B	C	D	E	F
Before campaign :		53	28	31	48	50	42
After campaign :		58	29	30	55	56	45

Can the campaign be judged to be a success? Test at 5 per cent level of significance.

[Jammu Univ., MCom, 1997]

- 8.60** As a controller of budget you are presented with the following data for budget variances (in Rs 000's)

Department	Budgeted sales	Actual sales
A	1000	900
B	850	880
C	720	650
D	1060	860
E	750	820
F	900	1000
G	620	700
H	600	540
I	700	690
J	700	730
K	950	850
L	1100	1080

Is there any reason that achievements against budgets are slipping? Take $\alpha = 0.05$ level of significance.

- 8.61** A certain medicine given to each of 12 patients resulted in the following increase in blood pressure: 2, 5, 8, -1, 3, 0, -2, 1, 5, 0, 4, 6. Can it be concluded that the medicine will, in general, be accompanied by an increase in blood pressure?

- 8.62** In a survey of buying habits, 400 women shoppers are chosen at random in a shopping mall A. Their average monthly food expenditure is Rs. 2500 with a standard deviation of Rs 400, for another group of 400 women shoppers chosen at random in shopping mall B, located in another area of the same city, the average monthly food expenditure is Rs 2200 with a standard deviation of Rs 550. Do the data present sufficient evidence to indicate that the average monthly food expenditure of the two populations of women shoppers are equal. Test at the 1 per cent level of significance.

- 8.63** The average monthly earnings for a women in managerial and professional positions is Rs 16,700. Do men in the same positions have average monthly earnings that are higher than those for women? A random sample of $n=40$ men in managerial and professional positions showed $\bar{x} = \text{Rs } 17,250$ and $s = \text{Rs } 2346$. Test the appropriate hypothesis using $\alpha = 0.01$.

- 8.64** The director (finance) of a company collected the following sample information to compare the daily travel expenses for the sales staff and the audit staff:

Sales (Rs) : 262 270 292 230 272 284

Audit (Rs): 260 204 258 286 298 240 278

At the $\alpha = 0.10$ significance level, can he conclude that the mean daily expenses are greater for the sales staff.

- 8.65** The owner of the weight lifting centre claims that by taking a special vitamin, a weight lifter can increase his strength. Ten student athletes are randomly selected and given a test of strength using the standard bench press. After two-weeks of regular training, supplemented with the vitamin, they are tested again.

The results are shown below:

Student: 1 2 3 4 5 6 7 8 9 10

Before : 190 250 345 210 114 126 186 116 196 125

After : 196 240 345 212 113 129 189 115 194 124

At the $\alpha = 0.01$ significance level, can we conclude that the special vitamin increased the strength of the student athletes?

- 8.66** The manufacturer of the motorcycle advertises that the motorcycle will average 87 kms per litre on long trips. The kms on eight long trips were 88, 82, 81, 87, 80, 78, 79 and 89. At the $\alpha = 0.05$ significance level, is the mean kilometer less than the advertised 87 kilometers per litre.
- 8.67** The variability in the amount of impurities present in a batch of chemical used for a particular process depends on the length of time the process is in operation. A manufacturer using two production lines 1 and 2 has

made a slight adjustment to line 2, hoping to reduce the variability as well as the average amount of impurities in the chemical. Samples of $n_1 = 25$ and $n_2 = 25$ measurements from two batches yield following means and variances: $\bar{x}_1 = 3.2$, $s_1^2 = 1.04$ and $\bar{x}_2 = 3.0$, $s_2^2 = 0.51$. Do the data present sufficient evidence to indicate that the process variability is less for line 2?

- 8.68** A media research group conducted a study of the radio listening habits of men and women. It was discovered that the mean listening time for men is 35 minutes per day with a standard deviation of 10 minutes in a sample of 10 men studied. The mean listening time for 12 women studied was also 35 minutes with a standard deviation of 12 minutes. Can it be concluded that there is a difference in the variation in the number of minutes men and women listen to the radio at $\alpha = 0.10$ significance level?

Hints and Answers

- 8.42** Let H_0 : No significant difference between the and hypothesized population means

Given, $\bar{x} = 33$, $s = \sqrt{100} = 10$, $n = 25$, and $\mu = 27$.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{33 - 27}{10/\sqrt{25}} = 3$$

Since $t_{\text{cal}} = 3$ is more than its critical value $t = 2.064$ at $\alpha/2 = 0.025$ and $df = 24$, the H_0 is rejected.

- 8.43** Let H_0 : No significant difference between the claim of the educator and the sample results

Given, $n = 150$, $\bar{x} = 111.2$, $s = 7.2$ and $\mu = 110$

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{111.2 - 110}{7.2/\sqrt{150}} = 2.04$$

Since $z_{\text{cal}} = 2.04$ is less than its critical value $z_{\alpha} = 2.58$ at $\alpha = 0.01$, the H_0 is accepted.

- 8.44** Let H_0 : $\mu = 65$ and H_1 : $\mu \neq 65$

x	$d = x - 65$	d^2
66	1	1
65	0	0
69	4	16
70	5	25
69	4	16
71	6	36
70	6	36
63	-2	4
64	-1	1
68	3	9
	25	133

$$\bar{x} = A + \frac{\Sigma d}{n} = 65 + \frac{25}{10} = 67.5 ;$$

$$s = \sqrt{\frac{\Sigma d^2}{n-1} - \frac{(\Sigma d)^2}{n(n-1)}} = \sqrt{\frac{133}{9} - \frac{(25)^2}{10 \times 9}} = 2.798$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{67.5 - 65}{2.798/\sqrt{10}} = 2.81$$

Since $t_{\text{cal}} = 2.81$ is more than its critical value $t_{\alpha/2} = 1.833$ at $\alpha/2 = 0.025$ and $df = 9$, the H_0 is rejected.

- 8.45** Population of bad apples in the given sample,
 $p = 50/500 = 0.1$; $q = 0.9$

$$\text{Standard error, } \sigma_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.1 \times 0.9}{500}} = 0.013$$

Confidence limits: $p \pm 3 \sigma_p = 0.1 \pm 3(0.013)$;
 or $0.081 \leq p \leq 0.139$

- 8.46** (a) Let H_0 : $p = 0.46$ and H_1 : $p \neq 0.46$
 Given $p = 0.46$, $q = 0.54$, $n = 200$;

$$\sigma_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.46 \times 0.54}{200}} = 0.0352$$

Since 101 or more indicates a majority, as a continuous variable let us consider it as 100.5 and therefore the proportion is $100.5/200 = 0.5025$

$$\begin{aligned} P(x \geq 101) &= P\left[z \geq \frac{\bar{p} - p}{\sigma_p}\right] = P\left[z \geq \frac{0.5025 - 0.46}{0.0352}\right] \\ &= P[z \geq 1.21] \end{aligned}$$

Required probability = $0.5000 - 0.3869 = 0.1131$

$$(b) \sigma_p = \sqrt{\frac{0.46 \times 0.54}{1000}} = 0.0158$$

$$\begin{aligned} P(x \geq 100.5) &= P\left[z \geq \frac{\bar{p} - p}{\sigma_p}\right] = P\left[z \geq \frac{0.5025 - 0.46}{0.0158}\right] \\ &= P[z \geq 2.69] \end{aligned}$$

Required probability = $0.5000 - 0.4964 = 0.0036$

- 8.47** Given $n = 100$, $\bar{x} = 15$, $\sigma_2 = 25$;

$$\text{Standard error } \sigma_{\bar{x}} = \sigma/\sqrt{n} = 5/\sqrt{100} = 0.5$$

99 per cent confidence limits: $\bar{x} \pm 2.58 \sigma_{\bar{x}}$

$$15 \pm 2.58 (0.5); 3.71 \text{ to } 16.29$$

95 per cent Confidence limits: $\bar{x} \pm 1.96 \sigma_{\bar{x}}$
 $= 15 \pm 1.96(0.5); 14.02$ to 15.98

- 8.48 Let H_0 : The machine has not improved after overhauling, $H_0: p_1 = p_2$

Given, $p_1 = 20/400 = 0.050$, $p_2 = 10/300 = 0.033$

$$\begin{aligned} \therefore p &= \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} = \frac{20 + 10}{400 + 300} = 0.043; \\ q &= 1 - p = 0.957 \\ z &= \frac{p_1 - p_2}{\sigma_{p_1 - p_2}} = \frac{p_1 - p_2}{\sqrt{pq \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \\ &= \frac{0.050 - 0.033}{\sqrt{0.043 \times 0.957 \left(\frac{1}{400} + \frac{1}{300} \right)}} \\ &= \frac{0.050 - 0.033}{0.0155} = 1.096 \end{aligned}$$

Since $z_{\text{cal}} = 1.096$ is less than its critical value $z_\alpha = 1.96$ at $\alpha = 5$ per cent, the H_0 is accepted.

- 8.49 Let H_0 : No significant difference in the quality of two brands of bulbs, i.e. $H_0: \mu_1 = \mu_2$

Given $n_1 = 100$, $s_1 = 82$, $\bar{x}_1 = 1300$ and $n_2 = 100$, $s_2 = 93$, $\bar{x}_2 = 1288$

$$\begin{aligned} z &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{1300 - 1288}{\sqrt{\frac{(82)^2}{100} + \frac{(93)^2}{100}}} \\ &= \frac{12}{\sqrt{67.24 + 86.49}} = 0.968 \end{aligned}$$

Since $z_{\text{cal}} = 0.968$ is less than its critical value $z_\alpha = 1.96$ at $\alpha = 5$ per cent, the H_0 is accepted.

- 8.50 Let H_0 : The difference in the mean score of boys and girls is not significant,

Given $n_1 = 50$, $s_1 = 10$, $\bar{x}_1 = 75$ and $n_2 = 100$, $s_2 = 12$, $\bar{x}_2 = 70$

$$\begin{aligned} z &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{75 - 70}{\sqrt{\frac{(10)^2}{50} + \frac{(12)^2}{100}}} \\ &= \frac{5}{\sqrt{3.44}} = 2.695 \end{aligned}$$

Since $z_{\text{cal}} = 2.695$ is more than its critical value $z_\alpha = 2.58$ at $\alpha = 5$ per cent, the H_0 is rejected.

- 8.51 Let H_0 : There is no significant difference in the mean life of the two makes of bulbs,

Given $n_1 = 100$, $\bar{x}_1 = 1500$, $\sigma_1 = 50$ and $n_2 = 100$,

$\bar{x}_2 = 1550$, $\sigma_2 = 60$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{1500 - 1550}{\sqrt{\frac{(50)^2}{100} + \frac{(60)^2}{100}}} = -\frac{50}{\sqrt{100}} = -0.77$$

Since $z_{\text{cal}} = -0.77$ is more than its critical value $z_\alpha = -2.58$ at $\alpha = 1$ per cent, the H_0 is accepted.

- 8.52 Let H_0 : No significant difference in the efficacy of two drugs, that is, $H_0: \mu_1 = \mu_2$

Drug A: $\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{60}{5} = 12$; $\Sigma(x_1 - \bar{x}_1)^2 = 10$

Drug B: $\bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{77}{7} = 11$; $\Sigma(x_2 - \bar{x}_2)^2 = 44$

$$\therefore s = \sqrt{\frac{\Sigma (x_1 - \bar{x}_1)^2 + (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{10 + 44}{5 + 7 - 2}} = 2.324$$

$$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{12 - 11}{2.324} \sqrt{\frac{5 \times 7}{5 + 7}} \\ &= \frac{1}{2.324} \sqrt{\frac{35}{12}} = 0.735 \end{aligned}$$

Since $t_{\text{cal}} = 0.735$ is less than its critical value $t = 2.228$ at $\alpha = 5$ per cent and $df = n_1 + n_2 - 2 = 10$, the H_0 is accepted.

- 8.53 Let H_0 : No significant difference in the mean life of the two types of bulbs, i.e. $H_0: \mu_1 = \mu_2$

Type I: $n_1 = 8$, $\bar{x}_1 = 1234$, $s_1 = 36$; Type II: $n_2 = 7$, $\bar{x}_2 = 1136$, $s_2 = 40$

$$\therefore s = \sqrt{\frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{7 \times 36 + 6 \times 40}{8 + 7 - 2}} = 37.9$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{1234 - 1136}{37.9} \sqrt{\frac{8 \times 7}{8 + 7}}$$

$$= \frac{98}{37.9} \times 1.932 = 5$$

Since $t_{\text{cal}} = 5$ is more than its critical value $t_\alpha = 2.16$ at $\alpha = 5$ per cent and $df = n_1 + n_2 - 2 = 13$, the H_0 is rejected.

- 8.54 Let H_0 : Two means are statistically equal, that is, $H_0: \mu_1 = \mu_2$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{72.4 - 73.9}{\sqrt{\frac{(14.8)^2}{100} + \frac{(17.9)^2}{200}}} = -0.77$$

Since $z_{\text{cal}} = -0.77$ is more than its critical value $z_\alpha = -2.58$ at $\alpha = 1$ per cent, the H_0 is accepted.

- 8.55** Let H_0 : The difference in the mean strength of the two yarns is not significant,

Yarn A : $n_1 = 4$, $\bar{x}_1 = 52$, $s_1^2 = 42$; Yarn B: $n_2 = 9$, $\bar{x}_2 = 42$, $s_2^2 = 56$

$$\begin{aligned} \therefore s &= \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \\ &= \sqrt{\frac{3 \times 42 + 8 \times 56}{4+9-2}} = 7.22 \\ t &= \frac{\bar{x}_1 - \bar{x}_2}{s} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{52 - 42}{7.22} \sqrt{\frac{4 \times 9}{4+9}} \\ &= \frac{10}{7.22} \sqrt{\frac{36}{13}} = 2.3 \end{aligned}$$

Since $t_{\text{cal}} = 2.3$ is more than its critical value $t_\alpha = 1.796$ at $\alpha = 5$ per cent and $df = 11$, the H_0 is rejected.

- 8.56** Let H_0 : The difference in the average life of two makes of batteries is not significant,

Given $n_1 = 100$, $\bar{x}_1 = 36.5$, $\sigma_1 = 1.8$ and $n_2 = 80$, $\bar{x}_2 = 36.81$, $\sigma_2 = 1.5$

$$\begin{aligned} \therefore z &= \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \\ &= \frac{36.5 - 36.81}{\sqrt{\frac{(1.8)^2}{100} + \frac{(1.5)^2}{80}}} = \frac{-0.31}{0.246} = 1.22 \end{aligned}$$

Since $z_{\text{cal}} = 1.22$ is less than its critical value $z = 1.96$ at $\alpha = 5$ per cent, the H_0 is accepted.

- 8.57** Let H_0 : The difference in the average salary of the two divisions is not significant,

Given $n_1 = 12$, $\bar{x}_1 = 12,500$, $s_1 = 320$ and $n_2 = 10$, $\bar{x}_2 = 11,200$, $s_2 = 480$

$$\begin{aligned} \therefore s &= \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \\ &= \sqrt{\frac{11 \times (320)^2 + 9 \times (480)^2}{12+10-2}} = 400 \\ t &= \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \\ &= \frac{12,500 - 11,200}{400} \sqrt{\frac{12 \times 10}{12+10}} = 7.59 \end{aligned}$$

Since $t_{\text{cal}} = 7.59$ is more than its critical value $t_\alpha = 2.228$ at $\alpha = 5$ per cent and $df = n_1 + n_2 - 2 = 20$, the H_0 is rejected.

- 8.58** Let H_0 : Overhauling has not improved the performance of the machine.

Given $p_1 = 10/200 = 0.05$; $p_2 = 4/100 = 0.04$;

$$\begin{aligned} \therefore p &= \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{10+4}{200+100} = 0.047; \\ q &= 1 - p = 0.953 \end{aligned}$$

$$\begin{aligned} \sigma_p &= \sqrt{pq \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \\ &= \sqrt{0.047 \times 0.953 \left(\frac{1}{200} + \frac{1}{100} \right)} = 0.027 \\ z &= \frac{p_1 - p_2}{\sigma_p} = \frac{0.05 - 0.04}{0.027} = 0.370 \end{aligned}$$

Since $z_{\text{cal}} = 0.370$ is less than its critical value $z = 1.96$ at $\alpha = 5$ per cent, the H_0 is accepted.

- 8.59** Let H_0 : The promotional campaign has not been successful, i.e. $\mu_1 = \mu_2$

d : 5 1 -1 7 6 3 = 21

d^2 : 25 1 1 49 36 9 = 121

$$\bar{d} = \frac{\sum d}{n} = \frac{21}{6} = 3.5;$$

$$s = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}} = \sqrt{\frac{121}{5} - \frac{(21)^2}{5 \times 6}} = 3.08$$

$$t = \frac{\bar{d}}{s/\sqrt{n}} = \frac{3.5}{3.08/\sqrt{6}} = 2.78$$

Since $t_{\text{cal}} = 2.78$ is more than its critical value $t = 2.571$ at $\alpha = 5$ per cent and $df = n - 1 = 5$, the H_0 is rejected. Campaign is successful.

- 8.60** Let H_0 : Achievements in sales against budgets are slipping, i.e., $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 < \mu_2$

Department	d	d^2
A	-100	10,000
B	30	900
C	-70	4900
D	-200	40,000
E	70	4900
F	100	10,000
G	80	6400
H	-60	3600
I	-10	100
J	30	900
K	-100	10,000
L	-20	400
	-250	92,100

$$d = \frac{\sum d}{n} = \frac{250}{12} = 20.83;$$

$$s = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}} = \sqrt{\frac{92,100}{11} - \frac{(-250)^2}{12 \times 11}} = 88.87$$

$$\therefore t = \frac{\bar{d}}{s/\sqrt{n}} = \frac{20.83}{88.87/\sqrt{12}} = 0.812$$

Since $t_{\text{cal}} = 0.812$ is less than its critical value $t = 1.79$ at $\alpha = 5$ per cent and $df = 11$, the H_0 is accepted.

- 8.61** Let H_0 : Medicine is not accompanied by an increase in blood pressure, i.e.

$H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 < \mu_2$

d	d^2
2	4
5	25
8	64
-1	1
3	9
0	0
-2	4
1	1
5	25
0	0
4	16
6	36
31	185

$$\bar{d} = \frac{\sum d}{n} = \frac{31}{12} = 2.58;$$

$$s = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}} = \sqrt{\frac{185}{11} - \frac{(31)^2}{12 \times 11}} = 3.08$$

$$\therefore t = \frac{\bar{d}}{s/\sqrt{n}} = \frac{2.58}{3.08/\sqrt{12}} = 2.89$$

Since $t_{\text{cal}} = 2.89$ is more than its critical value $t = 1.796$ at $\alpha = 5$ per cent and $df = 11$, the H_0 is rejected. Medicine increases blood pressure.

- 8.62** $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 \neq \mu_2$; μ_1 and μ_2 = average monthly food expenditures of population 1 and 2 respectively.

Given: $n_1 = 400$, $\bar{x}_1 = 2500$, $s_1 = 400$ and $n_2 = 400$, $\bar{x}_2 = 2200$, $s_2 = 550$

$$\begin{aligned} z &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{2500 - 2200}{\sqrt{\frac{(400)^2}{400} + \frac{(550)^2}{400}}} \\ &= \frac{300 \times 20}{\sqrt{462500}} = \frac{6000}{680.07} = 8.822 \end{aligned}$$

Since $z_{\text{Cal}} (= 8.822) > z_{\alpha/2} (= 2.58)$ at $\alpha = 0.01$, H_0 is rejected. Hence, average monthly expenditure of two population of women shopper differ significantly.

- 8.63** $H_0: \mu = 16,700$ and $H_1: \mu > 16,700$; μ = average monthly salary for men

Given $n = 400$, $\bar{x} = 17,250$ and $s = 2346$.

$$\begin{aligned} z &= \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{17,250 - 16,700}{2346/\sqrt{400}} \\ &= \frac{550}{117.3} = 4.68 \end{aligned}$$

Since $z_{\text{cal}} (= 4.68) > z_{\alpha} = 2.58$, H_0 is rejected and hence we conclude that the average monthly earnings for men are significantly higher than for women.

- 8.64** $H_0: \mu_s \leq \mu_a$ and $H_1: \mu_s > \mu_a$; $df = 6 + 7 - 1 = 11$

$$\begin{aligned} s &= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{(6-1)(12.2)^2 + (7-1)(15.8)^2}{6+7-2} = 203.82 \end{aligned}$$

$$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{142.5 - 130.3}{\sqrt{203.82\left(\frac{1}{6} + \frac{1}{7}\right)}} \\ &= 1.536 \end{aligned}$$

Since $t_{\text{cal}} (= 1.536) > t_a (= 1.363)$ at $df = 11$, reject H_0 . The mean daily expenses are greater for sales staff.

- 8.65** $H_0: \mu_a \leq 0$ and $H_1: \mu_a > 0$. Calculate $\bar{d} = 0.10$ and $s_d = 4.28$. Apply test statistic

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{0.10}{4.28/\sqrt{10}} = 0.07.$$

Since $t_{\text{cal}} (= 0.07) < t_a (= 2.821)$ at $df = 9$, reject H_0 . Hence there has been no increase.

- 8.66** $H_0: \mu \geq 87$ and $H_1: \mu < 87$

Calculate $\bar{x} = \sum d/n = 664/8 = 83.0$,

$$\begin{aligned} s &= \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}} = \sqrt{\frac{55244}{8-1} - \frac{(664)^2}{8(8-1)}} \\ &= 4.342 \end{aligned}$$

Applying test statistic

$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{83 - 87}{4.342/\sqrt{8}} = -2.61$$

Since $t_{\text{cal}} (= -2.61) < t_a (= -1.895)$ at $df = 7$, reject H_0 . Hence, the kilometer is less than advertiser.

- 8.67** $H_0: \sigma_1^2$ and σ_2^2 and $H_1: \sigma_1^2 > \sigma_2^2$

Given $s_1^2 = 1.04$, $s_2^2 = 0.51$, $n_1 = n_2 = 25$. Applying the test statistic

$$f = s_1^2/s_2^2 = 1.04/0.51 = 2.04$$

Since $f_{\text{cal}} (= 2.04) > f_{\alpha} (= 1.70)$ at $df_1 = df_2 = 24$, reject H_0 . Hence variability of line 2 is less than that of line 1.

- 8.68** $H_0: \sigma_L^2 = \sigma_m^2$ and $H_1: \sigma_L^2 \neq \sigma_m^2$

Given $s_w^2 = 12$, $s_m^2 = 10$, $n_m = 10$, $n_w = 12$. Applying the test statistic

$$f = s_1^2/s_2^2 = (12)^2/(10)^2 = 1.44$$

Since $f_{\text{cal}} (= 1.44) < f_{\alpha} (= 3.10)$, accept H_0 . Hence there is no difference in the variations of the two populations.

There is no merit in equality unless it be equality with best.

—John Spalding

Analysis of Variance

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- understand how 'analysis of variance'(ANOVA) can be used to test for the equality of three or more population means.
- understand and use the terms like 'response variable', 'a factor' and 'a treatment' in the analysis of variance.
- learn how to summarize F-ratio in the form of an ANOVA table.

9.1 INTRODUCTION

In Chapter 8 we introduced hypothesis testing procedures to test the significance of differences between two sample means to understand whether the means of two populations are equal based upon two independent random samples. In all these cases the null hypothesis states that there is no significant difference among population mean, that is, $H_0 : \mu_1 = \mu_2$. However, there may be situations where more than two populations are involved and we need to test the significance of differences between three or more sample means. We also need to test the null hypothesis that three or more populations from which independent samples are drawn have equal (or homogeneous) means against the alternative hypothesis that population means are not all equal. Let $\mu_1, \mu_2, \dots, \mu_r$ be the mean value for population 1, 2, ..., r respectively. Then from sample data we intend to test the following hypotheses:

$$H_0 : \mu = \mu_2 = \dots = \mu_r \quad (9-1)$$

and

$$H_1 : \text{Not all } \mu_j \text{ are equal } (j = 1, 2, \dots, r)$$

In other words, the null and alternative hypotheses of population means imply that the null hypothesis should be rejected if any of the r sample means is different from the others.

For example, the production level in three shifts in a factory can be compared to answer of the questions such as: Is the production level higher/lower on any day of the week? Is Wednesday morning shift's production better/worse than any other shift? and so on. Production level can also be analysed using other days and shifts of the week in combination.

The following are a few examples involving more than two populations where it is necessary to conduct a comparative study to arrive at a statistical inference:

- Effectiveness of different promotional devices in term of sales
- Quality of a product produced by different manufacturers in terms of an attribute
- Production volume in different shifts in a factory
- Yield from plots of land due to varieties of seeds, fertilizers, and cultivation methods

Under certain circumstances we may not conduct repeated *t*-tests on pairs of the samples. This is because when many independent tests are carried out pairwise, the probability of the outcome being correct for the combined results is reduced greatly. Table 9.1 shows how the probability of being correct decreases when we intend to compare the average marks of 2, 3, and 10 students at the end of an examination at 95 per cent confidence level or 0.95 probability of being correct in our statistical inferences in this experiment.

Table 9.1: Calculations of Probability of being Correct

Number of Students, <i>n</i>	Number of Pairs	Confidence Level	Probability of Error
2	1	0.950	0.050
3	3	$(0.95)^3 = 0.857$	0.143
10	45	$(0.95)^{45} = 0.100$	0.900

It is clear from the calculations in Table 9.1 that as the size of student population or sample increases, the probability of error in statistical inference of population means increases. Under certain assumptions, a method known as **analysis of variance (ANOVA)** developed by R. A. Fisher is used to test the significance of the difference between several population means.

The following are few terms that will be used during discussion on analysis of variance:

- A *sampling plan or experimental design* is the way that a sample is selected from the population under study and determines the amount of information in the sample.
- An *experimental unit* is the object on which a measurement or measurements is taken. Any experimental conditions imposed on an experimental unit provides effect on the response.
- A *factor or criterion* is an independent variable whose values are controlled and varied by the researcher.
- A *level* is the intensity setting of a factor.
- A *treatment or population* is a specific combination of factor levels.
- The *response* is the dependent variable being measured by the researcher.

For example,

1. A tyre manufacturing company plans to conduct a tyre-quality study in which quality is the independent variable called *factor or criterion* and the *treatment levels or classifications* are low, medium and high quality. The *dependent (or response)* variable might be the number of kilometers driven before the tyre is rejected for use. A study of daily sales volumes may be taken by using a completely randomized design with demographic setting as the independent variable. A treatment levels or classifications would be inner-city stores, stores in metro-cities, stores in state capitals, stores in small towns, etc. The dependent variable would be sales in rupees.
2. For a production volume in three shifts in a factory, there are two variables—days of the week and the volume of production in each shift. If one of the objectives is to determine whether mean production volume is the same during days of the week, then the *dependance (or response) variable* of interest, is the mean production volume. The *variables* that are related to a response variable are called **factors**, that is, a day of the week is the *independent variable* and the value assumed by a factor in an experiment is called a **level**. The combinations of levels of the factors for which the response will be observed are called *treatments*, i.e. days of the week. These treatments define the populations or samples which are differentiated in terms of production volume and we may need to compare them with each other.

Analysis of variance: A statistical procedure for determining whether the means of several different populations are equal.

Factor: Another word for independent variable of interest that is controlled in the analysis of variance.

Factor level: A value at which the factor is controlled.

9.2 ANALYSIS OF VARIANCE APPROACH

The first step in the analysis of variance is to partition the total variation in the sample data into the following two component variations in such a way that it is possible to estimate the contribution of factors that may cause variation.

1. The amount of variation *among the sample means* or the variation attributable to the difference among sample means. This variation is either on account of difference in treatment or due to element of chance. This difference is denoted by SSC or SSTR.
2. The amount of variation *within the sample observations*. This difference is considered due to chance causes or experimental (random) errors. The difference in the values of various elements in a sample due to chance is called an estimate and is denoted by SSE.

The observations in the sample data may be classified according to *one factor* (criterion) or *two factors* (criteria). The classifications according to one factor and two factors are called *one-way classification* and *two-way classification*, respectively. The calculations for total variation and its components may be carried out in each of the two-types of classifications by (i) *direct method*, (ii) *short-cut method*, and (iii) *coding method*.

Assumptions for Analysis of Variance

The following assumptions are required for analysis of variance:

1. Each population under study is normally distributed with a mean μ_r that may or may not be equal but with equal variances σ_r^2 .
2. Each sample is drawn randomly and is independent of other samples.

9.3 TESTING EQUALITY OF POPULATION (TREATMENT) MEANS: ONE-WAY CLASSIFICATION

Many business applications involve experiments in which different populations (or groups) are classified with respect to only one attribute of interest such as (i) *percentage of marks* secured by students in a course, (ii) *flavour preference* of ice-cream by customers, (iii) *yield of crop* due to varieties of seeds, and so on. In all such cases observations in the sample data are classified into several groups based on a single attribute and is called **one-way classification** of sample data.

As mentioned before, for all theoretical purposes we refer populations (i.e., several groups classified based on single factor or criterion in a sample data) as treatments. We will study the effect of a factor (criterion) such as flavour preference on the dependent variable (i.e. sales) at different groups (i.e. variety of ice-creams). These groups are the treatments in this particular example.

Suppose our aim is to make inferences about r population means $\mu_1, \mu_2, \dots, \mu_r$ based on independent random samples of size n_1, n_2, \dots, n_r , from normal populations with a common variances σ^2 . That is, each of the normal population has same shape but their locations might be different. The null hypothesis to be tested is stated as:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_r \quad \leftarrow \text{Null hypothesis}$$

$$H_1 : \text{Not all } \mu_j \text{ } (j = 1, 2, \dots, r) \text{ are equal} \quad \leftarrow \text{Alternative hypothesis}$$

Let n_j = size of the j th sample ($j = 1, 2, \dots, r$)

n = total number of observations in all samples combined (i.e. $n = n_1 + n_2 + \dots + n_r$)

x_{ij} = the i th observation value within the sample from j th population

The observations values obtained for r independent samples based on one-criterion classification can be arranged as shown in Table 9.2.

One-way analysis of variance: Analysis of variance in which only one criterion (variable) is used to analyse the difference between more than two population means.

Table 9.2: One-Criterion Classification of Data

	Observations		Populations (Number of Samples)	
	1	2	...	r
1	x_{11}	x_{12}	...	x_{1r}
2	x_{21}	x_{22}	...	x_{2r}
\vdots	\vdots			
k	x_{k1}	x_{k2}	...	x_{kr}
Sum	T_1	T_2	...	$T_r = T$
A.M.	\bar{x}_1	\bar{x}_2	...	$\bar{x}_r = \bar{x}$

where

$$T_i = \sum_{i=1}^k x_{il} \quad T = \sum_{j=1}^r T_i$$

$$\bar{x}_i = \frac{1}{k} \sum_{i=1}^k x_{il} \quad \bar{x} = \frac{1}{rk} \sum_{j=1}^r \bar{x}_j = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^r x_{ij}$$

The values of \bar{x}_i are called sample means and \bar{x} is the grand mean of all observations (or measurements) in all the samples.

Since there are k rows and r columns in Table 9.2, therefore total number of observations are $rk = n$, provided each row has equal number of observations. But, if the number of observations in each row varies, then the total number of observations is $n_1 + n_2 + \dots + n_r = n$.

Illustration: Three brands A, B, and C of tyres were tested for durability. A sample of four tyres of each brand is subjected to the same test and the number of kilometres until wearout was noted for each brand of tyres. The data in thousand kilometres is given in Table 9.3.

Table 9.3: Example of Data in ANOVA

	Observations		Population (Number of Brands)
	A	B	
1	26	18	23
2	25	16	19
3	28	17	26
4	12	18	30
Sum	91	69	98
Sample size	94	94	94
Mean	22.75	17.25	24.50

Since the same number of observations is obtained from each brand of tyres (population), therefore the number of observations in the table is $n = rk = 3 \times 4 = 12$.

- The sample mean of each of three samples is given by

$$\bar{x}_1 = \frac{1}{4} \sum_{i=1}^4 x_{il} = \frac{1}{4} (91) = 22.75$$

$$\bar{x}_2 = \frac{1}{4} \sum_{i=1}^4 x_{i2} = \frac{1}{4} (69) = 17.25 \text{ and } \bar{x}_3 = 24.50$$

- The grand mean for all samples is

$$\bar{x} = \frac{1}{3} (\bar{x}_1 + \bar{x}_2 + \bar{x}_3) = \frac{1}{3} (22.75 + 17.25 + 24.50) = 21.50$$

9.3.1 Steps for Testing Null Hypothesis

Step 1: State the null and alternative hypotheses to test the equality of population means as:

$$\begin{aligned} H_0 : \mu_1 = \mu_2 = \dots = \mu_r &\quad \leftarrow \text{Null hypothesis} \\ H_1 : \text{Not all } \mu_j \text{ are equal } (j = 1, 2, \dots, r) &\quad \leftarrow \text{Alternative hypothesis} \end{aligned}$$

α = level of significance

Step 2: Calculate total variation If a single sample of size n is taken from the population, then estimate of the population variance based on the variance of sampling distribution of mean is given by

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{SS}{df}$$

The numerator in s^2 is called *sum of squares* of deviations of sample values about the sample mean \bar{x} and is denoted as SS. Consequently 'sum of squares' is a measure of variation. Thus when SS is divided by df , the result is often called the *mean square* which is an alternative term for sample variance.

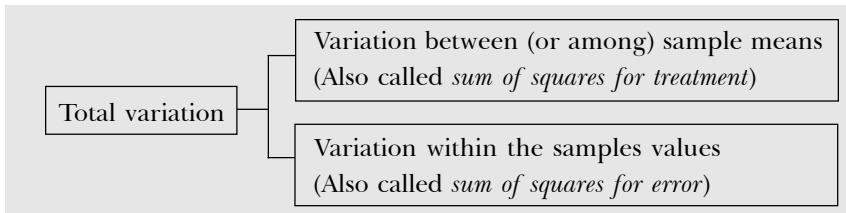
Total variation is represented by the '*sum of squares total*' (SST) and is equal to the sum of the squared differences between each sample value from the grand mean \bar{x}

$$SST = \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

where r = number of samples (or treatment levels)

n_j = size of j th sample

The total variation is divided into two parts as shown below:



Step 3: Calculate variation between sample means This is usually called the 'sum of squares between' and measures the variation between samples due to treatments. In statistical terms, variation between samples means is also called the *between-column variance*. The procedure is as follows:

- Calculate mean values $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_r$ of all r samples
- Calculate grand mean $\bar{x} = \frac{1}{r} (\bar{x}_1 + \bar{x}_2 + \dots + \bar{x}_r) = \frac{T}{n}$
where T = grand total of all observations.
 n = number of observations in all r samples.
- Calculate difference between the mean of each sample and the grand mean as $\bar{x}_1 - \bar{x}, \bar{x}_2 - \bar{x}, \dots, \bar{x}_r - \bar{x}$. Multiply each of these by the number of observations in the corresponding sample and add. The total gives the sum of the squared differences between the sample means in each group and is denoted by SSC or SSTR.

$$SSTR = \sum_{j=1}^r n_j (\bar{x}_j - \bar{x})^2$$

This sum is also called *sum of squares for treatment* (SSTR)

Step 4: Calculate variation within samples This is usually called the 'sum of squares within' and measures the difference within samples due to chance error. Such variation is also called *within sample variance*. The procedure is as follows:

- Calculate mean values $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_r$ of all r samples.

- (b) Calculate difference of each observation in r samples from the mean values of the respective samples.
- (c) Square all the differences obtained in Step (b) and find the total of these differences. The total gives the sum of the squares of differences within the samples and is denoted by SSE.

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_j)^2$$

This sum is also called the *sum of squares for error*, $SSE = SST - SSTR$.

Step 5: Calculate average variation between and within samples—mean squares Since r independent samples are being compared, therefore $r - 1$ degrees of freedom are associated with the sum of the squares among samples. As each of the r samples contributes $n_j - 1$ degrees of freedom for each independent sample within itself, therefore there are $n - r$ degrees of freedom associated with the sum of the squares within samples. Thus total degrees of freedom equal to the degrees of freedom associated with SSC (or SSTR) and SSE. That is

$$\begin{aligned} \text{Total } df &= \text{Between samples (treatments) } df + \text{Within samples (error) } df \\ n - 1 &= (r - 1) + (n - r) \end{aligned}$$

When these ‘sum of squares’ are divided by their associated degrees of freedom, we get the following variances or *mean square* terms:

$$MSTR = \frac{SSTR}{r - 1}; \quad MSE = \frac{SSE}{n - r}; \quad \text{and} \quad MST = \frac{SST}{n - 1}$$

It may be noted that the quantity $MSE = SSE/(n - r)$ is a pooled estimate of σ^2 (weighted average of all r sample variances whether H_0 is true or not)

Step 6: Apply F-test statistic with $r - 1$ degrees of freedom for the numerator and $n - r$ degrees of freedom for the denominator

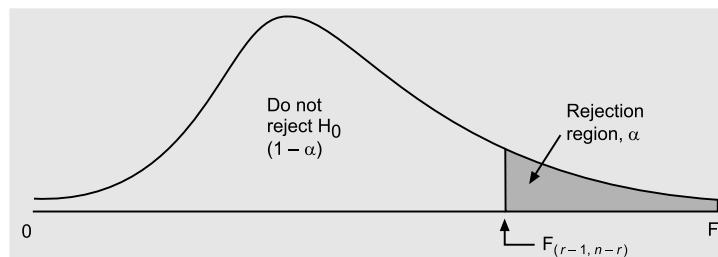
$$F = \frac{\sigma_{\text{between}}^2}{\sigma_{\text{within}}^2} = \frac{SSTR/(r - 1)}{SSE/(n - r)} = \frac{MSTR}{MSE}$$

Step 7: Make decision regarding null hypothesis If the calculated value of F-test statistic is more than its right tail critical value $F_{(r - 1, n - r)}$ at a given level of significance α and degrees of freedom $r - 1$ and $n - r$, then reject the null hypothesis. In other words, as shown in Fig. 9.1, the decision rule is:

- Reject H_0 if the calculated value of $F >$ its critical value $F_{\alpha(r - 1, n - r)}$
- Otherwise accept H_0

The F-distribution is a family of distributions, each identified by a pair of degrees of freedom. The first number refers to the number of degrees of freedom in the numerator of the F ratio, and the second refers to the number of degrees of freedom in the denominator. In the F table, columns represent the degrees of freedom for numerator and the rows represents the degrees of freedom for denominator.

Figure 9.1
Rejection Region for Null Hypothesis using ANOVA



If null hypothesis H_0 is true, then the variance in the sample means measured by $MSTR = SSTR/(r - 1)$ provides an unbiased estimate of σ^2 . But if H_0 is false, and population means are different, then MSTR is large as shown in Fig 9.2

Table 9.4 shows the general arrangement of the ANOVA table for one-factor analysis of variance.

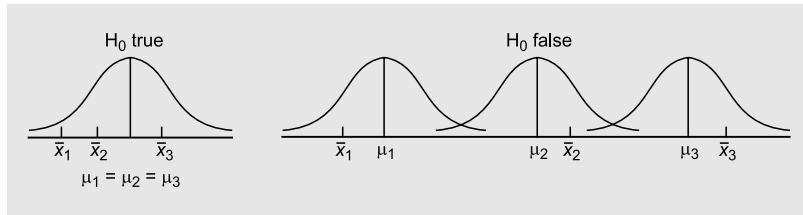


Figure 9.2
Sample Means Drawn from
Identical Populations

Table 9.4: ANOVA Summary Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	Test-Statistic or F-Value
• Between samples (Treatments)	SSTR	$r - 1$	$MSTR = \frac{SSTR}{r - 1}$	$F = \frac{MSTR}{MSE}$
• Within samples error	SSE	$n - r$	$MSE = \frac{SSE}{n - r}$	
Total	SST	$n - 1$		

Short-Cut Method The values of SSTR and SSE can be calculated by applying the following short-cut methods:

- Calculate the grand total of all observations in samples, T
- $$T = \sum x_1 + \sum x_2 + \dots + \sum x_r$$
- Calculate the correction factor $CF = \frac{T^2}{n}$; $n = n_1 + n_2 + \dots + n_r$
 - Find the sum of the squares of all observations in samples from each of r samples and subtract CF from this sum to obtain the total sum of the squares of deviations SST:

$$SST = (\sum x_1^2 + \sum x_2^2 + \dots + \sum x_r^2) - CF = \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}^2 - CF$$

$$SSTR = \frac{(\sum x_j)^2}{n_j} - CF$$

and

$$SSE = SST - SSTR$$

Coding Method Sometimes the method explained above takes a lot of computational time due to the magnitude of numerical values of observations. The coding method is based on the fact that the F-test statistic used in the analysis of variance is the ratio of variances without unit of measurement. Thus its values does not change if an appropriate constant value is either multiplied, divided, subtracted or added to each of the observations in the sample data. This adjustment reduces the magnitude of numerical values in the sample data and reduces computational time to calculate F value without any change.

Example 9.1: To test the significance of variation in the retail prices of a commodity in three principal cities, Mumbai, Kolkata, and Delhi, four shops were chosen at random in each city and the prices who lack confidence in their mathematical ability observed in rupees were as follows:

ANOVA table: A standard table used to summarize the analysis of variance calculations and results.

Mumbai :	16	8	12	14
Kolkata :	14	10	10	6
Delhi :	4	10	8	8

Do the data indicate that the price in the three cities are significantly different?

[Jammu Univ., M.Com, 1997]

Solution: Let us take the null hypothesis that there is no significant difference in the prices of a commodity in the three cities. Calculations for analysis of variance are as under.

Sample 1		Sample 2		Sample 3	
Mumbai		Kolkata		Delhi	
x_1	x_1^2	x_2	x_2^2	x_3	x_3^2
16	256	14	196	4	16
8	64	10	100	10	100
12	144	10	100	8	64
14	196	6	36	8	64
$\Sigma x_1 = 50$	$\Sigma x_1^2 = 660$	$\Sigma x_2 = 40$	$\Sigma x_2^2 = 432$	$\Sigma x_3 = 30$	$\Sigma x_3^2 = 244$

There are $r=3$ treatments (samples) with $n_1=4$, $n_2=4$, $n_3=4$, and $n=12$.

$$\begin{aligned} T &= \text{Sum of all the observations in the three samples} \\ &= \Sigma x_1 + \Sigma x_2 + \Sigma x_3 = 50 + 40 + 30 = 120 \end{aligned}$$

$$CF = \text{Correction factor} = \frac{T^2}{n} = \frac{(120)^2}{12} = 1200$$

$$\begin{aligned} SST &= \text{Total sum of the squares} \\ &= (\Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2) - CF = (660 + 432 + 244) - 1200 \\ &= 136 \end{aligned}$$

$$\begin{aligned} SSTR &= \text{Sum of squares between the samples} \\ &= \left(\frac{(\Sigma x_1)^2}{n_1} + \frac{(\Sigma x_2)^2}{n_2} + \frac{(\Sigma x_3)^2}{n_3} \right) - CF \\ &= \left\{ \frac{(50)^2}{4} + \frac{(40)^2}{4} + \frac{(30)^2}{4} \right\} - 1200 \\ &= \left\{ \frac{2500}{4} + \frac{1600}{4} + \frac{900}{4} \right\} - 1200 = \frac{5000}{4} - 1200 = 50 \end{aligned}$$

$$SSE = SST - SSTR = 136 - 50 = 86$$

$$\text{Degrees of freedom: } df_1 = r - 1 = 3 - 1 = 2 \quad \text{and} \quad df_2 = n - r = 12 - 3 = 9$$

$$\text{Thus MSTR} = \frac{SSTR}{df_1} = \frac{50}{2} = 25 \quad \text{and} \quad MSE = \frac{SSE}{df_2} = \frac{86}{9} = 9.55$$

Table 9.5: ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	Test-Statistic
• Between samples	50	2	25	$F = \frac{25}{9.55} = 2.617$
• Within samples	86	9	9.55	
Total	136	11		

The table value of F for $df_1 = 2$, $df_2 = 9$, and $\alpha = 5$ per cent level of significance is 4.26. Since calculated value of F is less than its critical (or table) value, the null hypothesis is accepted. Hence we conclude that the prices of a commodity in three cities have no significant difference.

Example 9.2: A study investigated the perception of corporate ethical values among individuals specializing in marketing. Using $\alpha = 0.05$ and the following data (higher scores indicate higher ethical values), test for significant differences in perception among three groups.

<i>Marketing Manager</i>	<i>Marketing Research</i>	<i>Advertising</i>
6	5	6
5	5	7
4	4	6
5	4	5
6	5	6
4	4	6

Solution: Let us assume the null hypothesis that there is no significant difference in ethical values among individuals specializing in marketing. Calculations for analysis of variance are as under:

<i>Marketing Manager</i>		<i>Marketing Research</i>		<i>Advertising</i>	
x_1	x_1^2	x_2	x_2^2	x_3	x_3^2
6	36	5	25	6	36
5	25	5	25	7	49
4	16	4	16	6	36
5	25	4	16	5	25
6	36	5	25	6	36
4	16	4	16	6	36
30	154	27	123	36	218

There are $r = 3$ treatments (samples) with $n_1 = n_2 = n_3 = 6$ and $n = 18$.

T = Sum of all the observations in three samples

$$= \sum x_1 + \sum x_2 + \sum x_3 = 30 + 27 + 36 = 93$$

$$\text{CF} = \text{Correction factor} = \frac{T^2}{n} = \frac{(93)^2}{18} = 480.50$$

SST = Total sum of the squares

$$= (\sum x_1^2 + \sum x_2^2 + \sum x_3^2) - \text{CF} = (154 + 123 + 218) - 480.50 \\ = 14.50$$

SSTR = Sum of squares between the samples

$$= \left(\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} \right) - \text{CF}$$

$$= \left\{ \frac{(30)^2}{6} + \frac{(27)^2}{6} + \frac{(36)^2}{6} \right\} - 480.50$$

$$= \left(\frac{900}{6} + \frac{729}{6} + \frac{1296}{6} \right) - 480.50$$

$$= (150 + 121.5 + 216) - 480.50 = 7$$

$$\text{SSE} = \text{SST} - \text{SSTR} = 14.50 - 7 = 7.50$$

Degrees of freedom: $df_1 = r - 1 = 3 - 1 = 2$ and $df_2 = n - r = 18 - 3 = 15$

Thus $\text{MSTR} = \frac{\text{SSTR}}{df_1} = \frac{7}{2} = 3.5$ and $\text{MSE} = \frac{\text{SSE}}{df_2} = \frac{7.50}{15} = 0.5$

Table 9.6: ANOVA Table

<i>Source of Variation</i>	<i>Sum of Squares</i>	<i>Degrees of Freedom</i>	<i>Mean Squares</i>	<i>Test-Statistic</i>
• Between samples	7	2	3.5	$F = \frac{3.5}{0.5}$
• Within samples	7.5	15	0.5	= 7
Total	14.5	17		

The table value of F for $df_1 = 2$, $df_2 = 15$, and $\alpha = 0.05$ is 3.68. Since calculated value of F=7 is more than its table value, the null hypothesis is rejected. Hence we conclude that there is significant difference in ethical values among individuals specializing in marketing.

Example 9.3: As head of the department of a consumer's research organization, you have the responsibility for testing and comparing lifetimes of four brands of electric bulbs. Suppose you test the life-time of three electric bulbs of each of the four brands. The data are shown below, each entry representing the lifetime of an electric bulb, measured in hundreds of hours:

<i>Brand</i>			
<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
20	25	24	23
19	23	20	20
21	21	22	20

Can we infer that the mean lifetimes of the four brands of electric bulbs are equal?
[Roorkee Univ., MBA, 2000]

Solution: Let us take the null hypothesis that the mean lifetime of the four brands of electric bulbs is equal.

Substracting a common figure 20 from each observation. The calculations with code data are as under:

<i>Sample 1(A)</i>		<i>Sample 2(B)</i>		<i>Sample 3(C)</i>		<i>Sample 4(D)</i>	
x_1	x_1^2	x_2	x_2^2	x_3	x_3^2	x_4	x_4^2
0	0	5	25	4	16	3	9
-1	1	3	9	0	0	0	0
1	1	1	1	2	4	0	0
0	2	9	35	6	20	3	9

$$\begin{aligned} T &= \text{Sum of all the observations in four samples} \\ &= \Sigma x_1 + \Sigma x_2 + \Sigma x_3 + \Sigma x_4 = 0 + 9 + 6 + 3 = 18 \end{aligned}$$

$$CF = \text{Correction factor} = \frac{T^2}{n} = \frac{(18)^2}{12} = 27$$

$$\begin{aligned} SST &= \text{Total sum of the squares} \\ &= (\Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 + \Sigma x_4^2) - CF \\ &= (2 + 35 + 20 + 9) - 27 = 39 \end{aligned}$$

SSTR = Sum of squares between the samples

$$= \left\{ \frac{(\Sigma x_1)^2}{n_1} + \frac{(\Sigma x_2)^2}{n_2} + \frac{(\Sigma x_3)^2}{n_3} + \frac{(\Sigma x_4)^2}{n_4} \right\} - CF$$

$$= \left\{ 0 + \frac{(9)^2}{3} + \frac{(6)^2}{3} + \frac{(3)^2}{3} \right\} - 18 = (0 + 27 + 12 + 3) - 27 = 15$$

$$SSE = SST - SSTR = 39 - 15 = 24$$

Degrees of freedom: $df_1 = r - 1 = 4 - 1 = 3$, $df_2 = n - r = 12 - 4 = 8$. Thus

$$\text{MSTR} = \frac{\text{SSTR}}{df_1} = \frac{15}{3} \quad \text{and} \quad \text{MSE} = \frac{\text{SSE}}{df_2} = \frac{24}{8} = 3$$

Table 9.7: ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	Test Statistic
• Between samples	15	3	5	$F = \frac{5}{3} = 1.67$
• Within samples	24	8	3	
Total	39	11		

The table value of F for $df_1 = 3$, $df_2 = 8$, and $\alpha = 0.05$ is 4.07. Since the calculated value of $F = 1.67$ is less than its table value, the null hypothesis is accepted. Hence we conclude that the difference in the mean lifetime of four brands of bulbs is not significant and we infer that the average lifetime of the four brands of bulbs is equal.

9.4 INFERENCES ABOUT POPULATION (TREATMENT) MEANS

When null hypothesis H_0 is rejected, it implies that all population means are not equal. However, we may not be satisfied with this conclusion and may want to know which population means differ. The answer to this question comes from the construction of confidence intervals using the small sample procedures, based on t-distribution.

For a single population mean, μ the confidence interval is given by

$$\bar{x} \pm t_{\alpha/2}(s/\sqrt{n})$$

where \bar{x} is the sample mean from a population. Similarly, confidence interval for the difference between two population means μ_1 and μ_2 is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

where \bar{x}_1 and \bar{x}_2 = mean of sample population 1 and 2 respectively;

n_1 and n_2 = number of observations in sample 1 and 2 respectively.

To use these confidence intervals, we need to know

- How to calculate s or s^2 , which is the best estimate of the common sample variance?
- How many degrees of freedom are used for the critical value of t -test statistic?

To answer these questions, recall that in the analysis of variance, we assume that the population variances are equal for all populations. This common value is the **mean square error** $MSE = SSE/(n - r)$ which provides an unbiased estimate of σ^2 , regardless of test or estimation used. We use, $s^2 = MSE$ or $s = \sqrt{MSE} = \sqrt{SSE/(n - r)}$ with $df = (n - r)$ and $t_{\alpha/2}$ at specified level of significance α to estimate σ^2 , where $n = n_1 + n_2 + \dots + n_r$.

Mean square error

(MSE): The mean of the squared errors used to judge the quality of a set of errors.

Illustration

From Example 9.1 we know that, $\bar{x}_1 = 5$; $\bar{x}_3 = 6$, $n = n_1 + n_2 + n_3 = 18$, $s^2 = MSE = 0.5$ and $\alpha = 0.05$ level of significance. So the confidence interval is computed as:

$$\begin{aligned} (\bar{x}_1 - \bar{x}_3) \pm t_{\alpha/2} \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} &= (5 - 6) \pm 2.131 \sqrt{0.5 \left(\frac{1}{6} + \frac{1}{6} \right)} \\ &= -1 \pm 1.225 = -2.225 \text{ and } 0.225 \end{aligned}$$

Since zero is included in this interval, we may conclude that there is no significant difference in the selected population means. That is, there is no difference between ethical values of marketing and advertising managers.

Remark: If the end points of the confidence interval have the same sign, then we may conclude that there is a significant difference between the selected population means.

Self-Practice Problems 9A

- 9.1** Kerala Traders Co. Ltd., 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 selling ability as measured by the average size of 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 :	300	400	300	500	0
B :	600	300	300	400	—
C :	700	300	400	600	500

Perform the analysis of variance and draw your conclusions.

[Madras Univ., MCom, 1996; Madurai Univ., MCom, 1996]

- 9.2** There are three main brands of a certain powder. A sample of 120 packets sold is examined and found to be allocated among four groups A, B, C, and D, and brands I, II and III, as shown below:

Brand	Group			
	A	B	C	D
I	10	14	18	15
II	15	18	13	16
III	18	19	11	13

Is there any significant difference in brand preferences?

- 9.3** An agriculture research organization wants to study the effect of four types of fertilizers on the yield of a crop. It divided the entire field into 24 plots of land and used fertilizer at random in 6 plots of land. Part of the calculations are shown below:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F-Test Statistic
• Fertilizers	2940	3	—	5.99
• Within groups	—	—	—	
Total	6212	24	—	

- (a) Fill in the blanks in the ANOVA table.
 (b) Test at $\alpha = 0.05$, whether the fertilizers differ significantly

- 9.4** A manufacturing company has purchased three new machines of different makes and wishes to determine whether one of them is faster than the others in

producing a certain output. Five hourly production figures are observed at random from each machine and the results are given below:

Observations	A_1	A_2	A_3
1	25	31	24
2	30	39	30
3	36	38	28
4	38	42	25
5	31	35	28

Use analysis of variance and determine whether the machines are significantly different in their mean speed.

- 9.5** The following figures related to the number of units of a product sold in five different areas by four salesmen:

Area	Number of units			
	A	B	C	D
1	80	100	95	70
2	82	110	90	75
3	88	105	100	82
4	85	115	105	88
5	75	90	80	65

Is there a significant difference in the efficiency of these salesmen?

[Osmania Univ., MBA, 1998]

- 9.6** Four machines A, B, C, and D are used to produce a certain kind of cotton fabric. Samples of size 4 with each unit as 100 square metres are selected from the outputs of the machines at random, and the number of flowers in each 100 square metres are counted, with the following results:

A	Machines		
	B	C	D
8	6	14	20
9	8	12	22
11	10	18	25
12	4	9	23

Do you think that there is significant difference in the performance of the four machines?

[Kumaon Univ., MBA, 1998]

Hints and Answers

- 9.1** Let H_0 : No difference in average sales of three salesmen.

Divide each observation by 100 and use the code data for analysis of variance.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Test Statistic
• Between Samples	10	2	5	$F = \frac{5}{2.73} = 1.83$
• Within Samples	30	11	2.73	
Total	40	13		

Since the calculated value of $F = 1.83$ is less than its table value $F = 3.98$ at $df_1 = 2$, $df_2 = 11$, and $\alpha = 0.05$, the null hypothesis is accepted.

- 9.2 Let H_0 : There is no significant difference in brand preference.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Test Statistic
• Between Samples	168.5	2	84.25	$F = \frac{84.25}{22.83} = 3.69$
• Within Samples	205.5	9	22.83	
Total	374.0	11		

Since calculated value of $F = 3.69$ is less than its table value $F = 4.26$ at $df_1 = 2$, $df_2 = 9$, and $\alpha = 0.05$, the null hypothesis is accepted.

- 9.3 Given total number of observations, $n = 24$; Number of samples, $r = 4$

$df = n - 1 = 24 - 1 = 23$ (For between the groups-fertilizers)

$df_1 = r - 1 = 4 - 1 = 3$; $df_2 = n - r = 24 - 4 = 20$
(For within the groups)

$$SSTR = 2940;$$

$$SSE = SST - SSB = 6212 - 2940 = 3272$$

$$MSTR = \frac{SSTR}{df_1} = \frac{2940}{3} = 980;$$

$$MSE = \frac{SSE}{df_2} = \frac{3272}{20} = 163.6$$

$$\therefore F = \frac{MSTR}{MSE} = \frac{980}{163.6} = 5.99$$

Since the calculated value of $F = 5.99$ is more than its table value $F = 3.10$ at $df_1 = 3$, $df_2 = 20$, and $\alpha = 0.05$, the null hypothesis is rejected.

- 9.4 Let H_0 : Machines are not significantly different in their mean speed.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Test Statistic
• Between Samples	250	2	125	$F = \frac{125}{16.66} = 7.50$
• Within Samples	200	12	16.66	
Total	450	14		

Since the calculated value of $F = 7.50$ is more than its table value $F = 3.89$ at $df_1 = 2$, $df_2 = 12$, and $\alpha = 0.05$, the null hypothesis is rejected.

- 9.5 Let H_0 : No significant difference in the performance of four salesmen.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Test Statistic
• Between Samples	2340	3	780	$F = \frac{780}{73.5} = 7.50$
• Within Samples	1176	16	73.5	
Total	3516	19		

Since the calculated value of $F = 10.61$ is greater than its table value $F = 3.24$ at $df_1 = 3$, $df_2 = 16$, and $\alpha = 0.05$, the null hypothesis is rejected.

- 9.6 Let H_0 : Machines do not differ significantly in performance.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Test Statistic
• Between Samples	540.69	3	180.23	$F = \frac{180.23}{7.15} = 7.50$
• Within Samples	85.75	12	7.15	
Total	626.44	15		

Since the calculated value of $F = 25.207$ is more than its table value $F = 5.95$ at $df_1 = 3$, $df_2 = 12$, and $\alpha = 0.05$, the null hypothesis is rejected.

9.5 TESTING EQUALITY OF POPULATION (TREATMENT) MEANS: TWO-WAY CLASSIFICATION

In one-way ANOVA, the partitioning of the total variation in the sample data is done into two components: (i) Variation among the samples due to different samples (or treatments) and (ii) Variation within the samples due to random error. However, there might be a possibility that some of the variation left in the random error from one-way analysis of variation was not due to random error or chance but due to some other

measurable factor. For instance, in Example 9.1 we might feel that part of the variation in price was due to the inability in data collection or condensation of data. If so, this accountable variation was deliberately included in the sum of squares for error (SSE) and therefore caused the mean sum of squares for error (MSE) to be little large. Consequently, F-Value would then be small and responsible for the rejection of null hypothesis.

Two-way analysis of variance: Analysis of variance in which two criteria (or variables) are used to analyse the difference between more than two population means.

Blocking: The removal of a source of variation from the error term in the analysis of variance.

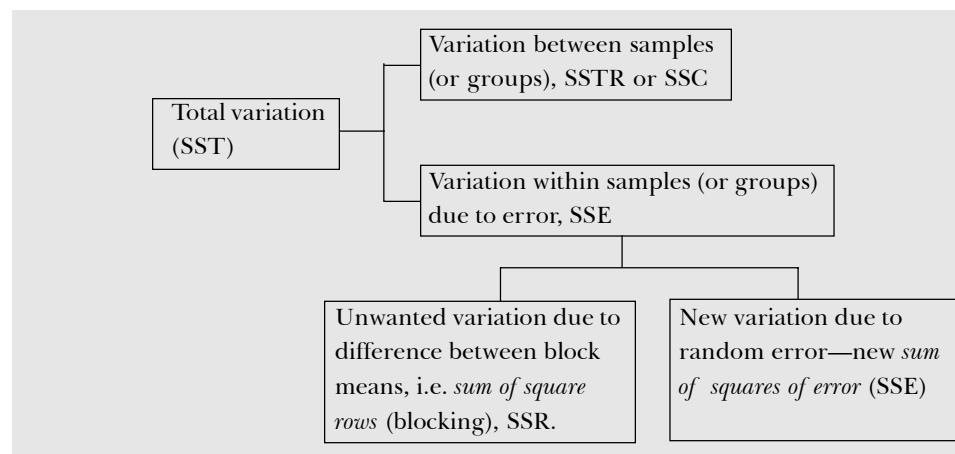
The **two-way analysis of variance** can be used to

- explore one criterion (or factor) of interest to partition the sample data so as to remove the unaccountable variation, and arriving at a true conclusion.
- investigate two criteria (factors) of interest for testing the difference between sample means.
- consider any interaction between two variables.

In two-way analysis of variance we are introducing another term called '*blocking variable*' to remove the undesirable accountable variation. A block variable is the variable that the researcher wants to control but is not the treatment variable of interest. The term '**blocking**' refers to block of land and comes from agricultural origin. The 'block' of land might make some difference in the study of growth pattern of varieties of seeds for a given type of land. R. A. Fisher designated several different plots of land as blocks, which he controlled as a second variable. Each of the seed varieties were planted on each of the blocks. The main aim of his study was to compare the seed varieties (independent variable). He only wanted to control the difference in plots of land (blocking variable). For instance, in Example 9.1, each set of three prices in three cities under a given condition would constitute a 'block' of sample data. 'Blocking' is an extension of the idea of pairing observations in hypothesis testing. Blocking provides the opportunity for one-to-one comparison of prices, where any observed difference cannot be due to difference among blocking variables.

To ensure a right conclusion to be reached, each sample data (group) should be measured under the same conditions by removing variations due to these conditions by the use of a blocking factor.

The partitioning of total variation in the sample data is shown below:



The general ANOVA table for c samples (columns), r blocks, and number of observations n is shown in Table 9.8.

Table 9.8: General ANOVA Table for Two-way Classification

Source of Variation	Sum of Square	Degrees of Freedom	Mean Square	Test Statistic
• Between columns	SSTR	$c - 1$	$MSTR = SSTR/(c - 1)$	$F_{\text{treatment}} = MSTR/MSE$
• Between rows	SSR	$r - 1$	$MSR = SSR/(r - 1)$	$F_{\text{blocks}} = MSR/MSE$
• Residual error	SSE	$(c - 1)(r - 1)$	$MSE = SSE/(c - 1)(r - 1)$	
Total	SST	$\frac{n - 1}{}$		

As stated above, total variation consists of three parts: (i) variation between columns, SSTR; (ii) variation between rows, SSR; and (iii) actual variation due to random error, SSE. That is

$$SST = SSTR + (SSR + SSE)$$

The degrees of freedom associated with SST are $cr - 1$, where c and r are the number of columns and rows, respectively

$$\text{Degrees of freedom between columns} = c - 1$$

$$\text{Degrees of freedom between rows} = r - 1$$

$$\text{Degrees of freedom for residual error} = (c - 1)(r - 1) = N - n - c + 1$$

The test-statistic F for analysis of variance is given by

$$F_{\text{treatment}} = MSTR/MSE; \quad MSTR > MSE \quad \text{or} \quad MSE/MSTR; \quad MSE > MSTR$$

$$F_{\text{blocks}} = MSR/MSE; \quad MSR > MSE \quad \text{or} \quad MSE/MSR; \quad MSE > MSR$$

Randomized block design:

A two-way analysis of variance designed to eliminate any assignable variation from the analysis.

Decision rule

- If $F_{\text{cal}} < F_{\text{table}}$, accept null hypothesis H_0
- Otherwise reject H_0

Example 9.5: The following table gives the number of refrigerators sold by 4 salesmen in three months May, June and July:

Month	Salesman			
	A	B	C	D
May	50	40	48	39
June	46	48	50	45
July	39	44	40	39

Is there a significant difference in the sales made by the four salesmen? Is there a significant difference in the sales made during different months?

[Delhi Univ., MCom, 1998]

Solution: Let us take the null hypothesis that there is no significant difference between sales made by the four salesmen during different months. The given data are coded by subtracting 40 from each observation. Calculations for a two-criteria—month and salesman—analysis of variance are shown in Table 9.9.

Table 9.9: Two-way ANOVA Table

Month	Salesman						Row Sum		
	$A(x_1)$	x_1^2	$B(x_2)$	x_2^2	$C(x_3)$	x_3^2	$D(x_4)$	x_4^2	
May	10	100	0	0	8	64	-1	1	17
June	6	36	8	64	10	100	5	25	29
July	-1	01	4	16	0	0	-1	1	2
Column sum	15	137	12	80	18	164	3	27	48

T = Sum of all observations in three samples of months = 48

$$CF = \text{Correction factor} = \frac{T^2}{n} = \frac{(48)^2}{12} = 192$$

SSTR = Sum of squares between salesmen (columns)

$$= \left\{ \frac{(15)^2}{3} + \frac{(12)^2}{3} + \frac{(18)^2}{3} + \frac{(3)^2}{3} \right\} - 192$$

$$= (75 + 48 + 108 + 3) - 192 = 42$$

SSR = Sum of squares between months (rows)

$$= \left\{ \frac{(17)^2}{4} + \frac{(29)^2}{4} + \frac{(2)^2}{4} \right\} - 192$$

$$= (72.25 + 210.25 + 1) - 192 = 91.5$$

$$\begin{aligned}
 SST &= \text{Total sum of squares} \\
 &= (\sum x_1^2 + \sum x_2^2 + \sum x_3^2 + \sum x_4^2) - CF \\
 &= (137 + 80 + 164 + 27) - 192 = 216 \\
 SSE &= SST - (SSC + SSR) = 216 - (42 + 91.5) = 82.5
 \end{aligned}$$

The total degrees of freedom are, $df = n - 1 = 12 - 1 = 11$. So

$$df_c = c - 1 = 4 - 1 = 3, df_r = r - 1 = 3 - 1 = 2; df = (c - 1)(r - 1) = 3 \times 2 = 6$$

Thus

$$MSTR = SSTR/(c - 1) = 42/3 = 14,$$

$$MSR = SSR/(r - 1) = 91.5/2 = 45.75$$

$$MSE = SSE/(c - 1)(r - 1) = 82.5/6 = 13.75$$

The ANOVA table is shown in Table 9.10.

Table 9.10: Two-way ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	Variance Ratio
• Between salesmen	42.0	3	14.00	$F_{\text{treatment}} = 14/13.75 = 1.018$
• Between months	91.5	2	45.75	$F_{\text{block}} = 45.75/13.75 = 3.327$
• Residual error	82.5	6	13.75	
Total	216	11		

(a) The table value of $F = 4.75$ for $df_1 = 3, df_2 = 6$, and $\alpha = 0.05$. Since the calculated value of $F_{\text{treatment}} = 1.018$ is less than its table value, the null hypothesis is accepted. Hence we conclude that sales made by the salesmen do not differ significantly

(b) The table value of $F = 5.14$ for $df_1 = 2, df_2 = 6$, and $\alpha = 0.05$. Since the calculated value of $F_{\text{block}} = 3.327$ is less than its table value, the null hypothesis is accepted. Hence we conclude that sales made during different months do not differ significantly.

Example 9.6: To study the performance of three detergents and three different water temperatures, the following 'whiteness' readings were obtained with specially designed equipment:

Water Temperature	Detergent A	Detergent B	Detergent C
Cold water	57	55	67
Warm water	49	52	68
Hot water	54	46	58

Perform a two-way analysis of variance, using 5 per cent level of significance.

[Osmania Univ., MBA, 1998]

Solution: Let us take the null hypothesis that there is no significant difference in the performance of three detergents due to water temperature and vice-versa. The data are coded by subtracting 50 from each observation. The data in coded form are in Table 9.11:

Table 9.11: Coded Data

Water Temperature	Detergents						Row Sum
	$A(x_1)$	x_1^2	$B(x_2)$	x_2^2	$C(x_3)$	x_3^2	
Cold water	+ 7	49	+ 5	25	+ 17	289	29
Warm water	- 1	01	+ 2	04	+ 18	324	19
Hot water	+ 4	16	- 4	16	+ 8	64	8
Column sum	10	66	3	45	43	677	56

$T = \text{Sum of all observations in three samples of detergents} = 56$

$$\text{CF} = \text{Correction factor} = \frac{T^2}{n} = \frac{(56)^2}{9} = 348.44$$

$SSTR = \text{Sum of squares between detergents (columns)}$

$$\begin{aligned} &= \left\{ \frac{(10)^2}{3} + \frac{(3)^2}{3} + \frac{(43)^2}{3} \right\} - \text{CF} \\ &= 33.33 + 3 + 616.33 - 348.44 = 304.22 \end{aligned}$$

$SSR = \text{Sum of squares between water temperature (rows)}$

$$\begin{aligned} &= \left\{ \frac{(29)^2}{3} + \frac{(19)^2}{3} + \frac{(8)^2}{3} \right\} - \text{CF} \\ &= (280.33 + 120.33 + 21.33) - 348.44 = 73.55 \end{aligned}$$

$SST = \text{Total sum of squares}$

$$= (\sum x_1^2 + \sum x_2^2 + \sum x_3^2) - \text{CF} = (66 + 45 + 677) - 348.44 = 439.56$$

$$SSE = SST - (SSC + SSR) = 439.56 - (304.22 + 73.55) = 61.79$$

Thus

$$MSTR = SSTR/(c - 1) = 304.22/2 = 152.11;$$

$$MSR = SSR/(r - 1) = 73.55/2 = 36.775$$

$$MSE = SSE/(c - 1)(r - 1) = 61.79/4 = 15.447$$

Table 9.12: Two-way ANOVA Table

Source of Variation	Sum of Squares	Degree of Freedom	Mean Squares	Variance Ratio
Between detergents (columns)	304.22	2	152.110	$F_{\text{treatment}} = 152.11/15.447 = 9.847$
Between temp. (rows)	73.55	2	36.775	$F_{\text{block}} = 36.775/15.447 = 2.380$
Residual error	61.79	4	15.447	
Total	439.56	8		

(a) Since calculated value of $F_{\text{treatment}} = 9.847$ at $df_1 = 2$, $df_2 = 4$, and, $\alpha = 0.05$ is greater than its table value $F = 6.94$, the null hypothesis is rejected. Hence we conclude that there is significant difference between the performance of the three detergents.

(b) Since the calculated value of $F_{\text{block}} = 2.380$ at $df_1 = 2$, $df_2 = 4$, and $\alpha = 0.05$ is less than its table value $F = 6.94$, the null hypothesis is accepted. Hence we conclude that the water temperature do not make a significant difference in the performance of the detergent.

Conceptual Questions 9A

- What are some of the criteria used in the selection of a particular hypothesis testing procedure?
- What are the major assumptions of ANOVA?
- Under what conditions should the one-way ANOVA F-test be selected to examine the possible difference in the means of independent populations?
- How is analysis of variance technique helpful in solving business problems? Illustrate your answer with suitable examples. [Kumaon Univ., MBA, 2000]
- Distinguish between one-way and two-way classifications to test the equality of population means.
- What is meant by the term analysis of variance? What types of problems are solved using ANOVA? Explain.
- Describe the procedure for performing the test of hypothesis in the analysis of variance. What is the basic assumption underlying this test?
- What is meant by the critical value used in the analysis of variance? How is it found?
- How is the F-distribution related to the student's t-distribution and the chi-square distribution? What important hypothesis can be tested by the F-distribution?

10. Discuss the components of total variation when samples are selected in blocks.
11. Define the terms treatment, error – ‘with in’ ‘between’ and the context in which these are used.
12. Explain the sum-of-square principle.
13. Explain how the total deviation is partitioned into the treatment deviation and the error deviation.
14. Does the quantity MSTR/MSE follow an F-distribution when the null hypothesis of ANOVA is false? Explain.

Self-Practice Problems 9B

- 9.7 A tea company appoints four salesmen A, B, C, and D, and observes their sales in three seasons—summer, winter and monsoon. The figures (in lakhs) are given in the following table:

Season	Salesman				Total
	A	B	C	D	
Summer	36	36	21	35	128
Winter	28	29	31	32	120
Monsoon	26	28	29	29	112
Totals	90	93	81	96	360

- (a) Do the salesmen significantly differ in performance?
 (b) Is there significant difference between the seasons?

[Calcutta Univ., MCom, 1996; Calcutta Univ., MCom, 1998]

- 9.8 Perform a two-way ANOVA on the data given below:

Plots of Land	Treatment			
	A	B	C	D
1	38	40	41	39
2	45	42	49	36
3	40	38	42	42

Use the coding method for subtracting 40 from the given numbers.
 [CA, May 1996]

- 9.9 The following data represent the production per day turned out by 5 different workers using 4 different types of machines:

Workers	Machine Type			
	A	B	C	D
1	44	38	47	36
2	46	40	52	43
3	34	36	44	32
4	43	38	46	33
5	38	42	49	39

- (a) Test whether the mean productivity is the same for the different machine types.
 (b) Test whether the 5 men differ with respect to mean productivity.
 [Madras Univ., MCom, 1997]

- 9.10 The following table gives the number of units of production per day turned out by four different types of machines:

Employees	Type of Machine			
	M ₁	M ₂	M ₃	M ₄
E ₁	40	36	45	30
E ₂	38	42	50	41
E ₃	36	30	48	35
E ₄	46	47	52	44

Using analysis of variance (a) test the hypothesis that the mean production is same for four machines and (b) test the hypothesis that the employees do not differ with respect to mean productivity.

[Osmania Univ., MCom, 1999]

- 9.11 In a certain factory, production can be accomplished by four different workers on five different types of machines. A sample study, in the context of a two-way design without repeated values, is being made with two fold objectives of examining whether the four workers differ with respect to mean productivity and whether the mean productivity is the same for the five different machines. The researcher involved in this study reports while analysing the gathered data as under:

- (a) Sum of squares for variance between machines = 35.2
 (b) Sum of squares for variance between workmen = 53.8
 (c) Sum of squares for total variance = 174.2
 Set up ANOVA table for the given information and draw the inference about variance at 5 per cent level of significance.

- 9.12 Apply the technique of analysis of variance of the following data showing the yields of 3 varieties of a crop each from 4 blocks, and test whether the average yields of the varieties are equal or not. Also test equality of the block means

Varieties	Blocks			
	I	II	III	IV
A	4	8	6	8
B	5	5	7	8
C	6	7	9	5

- 9.13 Three varieties of potato are planted each on four plots of land of the same size and type, each variety is treated with four different fertilizers. The yield in tonnes are as follows:

Fertilizer	Variety		
	V ₁	V ₂	V ₃
F ₁	164	172	174
F ₂	155	157	147
F ₃	159	166	158
F ₄	158	157	153

Perform an analysis of variance and show whether (a) there is any significant difference between the average yield of potatoes due to different fertilizers being used, and (b) there is any difference in the average yield of potatoes of different varieties.

Hints and Answers

- 9.7** Let H₀: No significant difference between sales by salesmen and that of seasons.

Decoding the data by subtracting 30 from each figure.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio
• Between salesmen (column)	42	3	14	$F_1 = \frac{22.67}{14} = 1.619$
• Between seasons (row)	32	2	16	$F_2 = \frac{22.67}{16} = 1.417$
• Residual error	136	6	22.67	
Total	210	11		

- Since F₁ = 1.619 < F_{0.05(6, 3)} = 4.76, accept null hypothesis.
- Since F₂ = 1.417 < F_{0.05(6, 2)} = 5.14, accept null hypothesis.

9.8

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio
• Between columns	42	3	14	$F_1 = \frac{14}{10.67} = 1.312$
• Between rows	26	2	13	$F_2 = \frac{13}{10.67} = 1.218$
• Residual error	64	6	10.67	
Total	132	11		

- (a) F₁ = 1.312 < F_{0.05(3, 6)} = 4.76, accept null hypothesis.
- (b) F₂ = 1.218 < F_{0.05(2, 6)} = 5.14, accept null hypothesis.

9.9 Let H₀ : (a) Mean productivity is same for all machines

(b) Men do not differ with respect to mean productivity

Decoding the data by subtracting 40 from each figure.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio
• Between machine	312.5	3	104.17	$F_1 = \frac{104.17}{10.72} = 9.72$
• Between employees	1266.0	3	88.67	$F_2 = \frac{88.67}{10.72} = 8.27$
• Residual error	96.5	9	10.72	
Total	675.0	15		

- (a) F_{0.05} = 3.86 at df₁ = 3 and df₂ = 9. Since the calculated value F₁ = 9.72 is more than its table value, reject the null hypothesis.
- (b) F_{0.05} = 3.86 at df₁ = 3 and df₂ = 9. Since the calculated value F₂ = 8.27 is more than its table value, reject the null hypothesis.

9.10 Let H₀: (a) Mean production does not differ for all machines

(b) Employees do not differ with respect to mean productivity

Decoding the data by subtracting 40 from each figure.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio
• Between machine	312.5	3	104.17	$F_1 = \frac{104.17}{10.72} = 9.72$
• Between employees	1266.0	3	88.67	$F_2 = \frac{88.67}{10.72} = 8.27$
• Residual error	96.5	9	10.72	
Total	675.0	15		

- (a) F_{0.05} = 3.86 at df₁ = 3 and df₂ = 9. Since the calculated value F₁ = 9.72 is more than its table value, reject the null hypothesis.
- (b) F_{0.05} = 3.86 at df₁ = 3 and df₂ = 9. Since the calculated value F₂ = 8.27 is more than its table value, reject the null hypothesis.

- 9.11** Let H_0 : (a) Workers do not differ with respect to their mean productivity
 (b) Mean productivity of all machines is the same

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio
• Between machine	35.2	4	8.8	$F_1 = \frac{8.8}{7.1} = 1.24$
• Between workmen	53.8	3	17.93	$F_2 = \frac{17.93}{7.1} = 2.53$
• Residual error	85.2	12	7.1	
• Total	<u>174.2</u>	<u>19</u>		

- (a) The calculated value of $F_1 = 1.24$ is less than its table value $F_{0.05} = 3.25$ at $df_1 = 4$ and $df_2 = 12$, hence the null hypothesis is accepted.
 (b) The calculated value of $F_2 = 2.53$ is less than its table value $F_{0.05} = 3.49$ at $df_1 = 3$ and $df_2 = 12$, hence the null hypothesis accepted.

- 9.12** Let H_0 : (a) Mean yields of the varieties are equal
 (b) Block means are equal

Decoding the data by subtracting 5 from each figure.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio
• Between blocks	9.67	3	3.22	$F_1 = \frac{3.22}{2.80} = 1.15$
• Between varieties	0.5	2	0.25	$F_2 = \frac{0.25}{0.25} = 11.20$
• Residual error	16.83	6	2.80	
Total	27.00	<u>12</u>		

- (a) Since the calculated value $F_1 = 1.15$ is less than its table value $F_{0.05} = 4.757$ at $df = (3, 6)$, the null hypothesis is accepted.

- (b) Since the calculated value $F_2 = 11.20$ is less than its table value $F_{0.05} = 19.33$ at $df = (6, 2)$, the null hypothesis is accepted.

- 9.13** Let H_0 : (a) No significant difference in the average yield of potatoes due to different fertilizers
 (b) No significant difference in the average yield of the three varieties of potatoes

Decoding the data by subtracting 158 from each figure.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio
• Between varieties	56	2	28	$F_1 = \frac{28}{18} = 1.55$
• Between fertilizers	498	3	166	$F_2 = \frac{166}{18} = 9.22$
• Residual error	108	6	18	
Total	<u>662</u>	<u>11</u>		

- (a) $F_{\text{cal}} = 1.55$ is less than its table value $F_{0.05} = 5.14$ at $df = (2, 6)$, the null hypothesis is accepted.

- (b) $F_{\text{cal}} = 9.22$ is more than its table value $F_{0.05} = 4.67$ at $df = (3, 6)$, the null hypothesis is rejected.

Formulae Used

1. One-way analysis of variance

- Grand sample mean

$$\bar{\bar{x}} = \frac{\sum_{i=1}^k \sum_{j=1}^r x_{ij}}{n}, n = n_1 + n_2 + \dots + n_r$$

- Correction factor CF = $\frac{T^2}{n}$

- Total sum of squares

$$SST = \sum_{i=1}^k \sum_{j=1}^r (x_{ij} - \bar{\bar{x}})^2 = \sum_i \sum_j x_{ij}^2 - CF$$

- Sum of squares of variations between samples due to treatment

$$SSTR = \sum_{j=1}^r n_j (\bar{x}_j - \bar{\bar{x}})^2 = \frac{1}{n_j} \sum_{j=1}^r x_j^2 - CF$$

- Sum of squares of variations within samples or error sum of squares

$$SSE = \sum_{i=1}^k \sum_{j=1}^r (x_{ij} - \bar{x}_i)^2 = SST - SSTR$$

- Mean square between samples due to treatments

$$MSTR = \frac{SSTR}{r-1}$$

- Mean square within samples due to error

$$MSE = \frac{SSE}{n-r}$$

- Test statistic for equality of k population means

$$F = \frac{MSTR}{MSE}$$

- Degrees of freedom

Total $df(n-1)$ = Treatment $df(r-1)$ + Random error $df(n-r)$

2. Two-way analysis of variance

- Total sum of squares

$$SST = \sum_{j=1}^r \sum_{i=1}^k x_{ij}^2 - n(\bar{x})^2$$

- Sum of squares of variances between columns due to treatments

$$SSTR = \sum_{j=1}^r (\bar{x}_j)^2 - n(\bar{\bar{x}})^2$$

- Sum of squares between rows due to blocks

$$SSR = \sum_{i=1}^k (\bar{x}_i)^2 - n(\bar{\bar{x}})^2$$

- Sum of squares due to error

$$SSE = SST - (SSTR + SSR)$$

- Degrees of freedom

$$df_c = c-1; df_r = (r-1)$$

$$\begin{aligned} df \text{ (residual error)} &= \text{Blocks } df + \text{Treatments } df \\ &= (r-1)(c-1) \end{aligned}$$

- Mean squares between columns due to treatment

$$MSTR = \frac{SSTR}{c-1}$$

- Mean square between rows due to blocks

$$MSR = \frac{SSR}{r-1}$$

- Mean square of residual error

$$MSE = \frac{SSE}{(c-1)(r-1)}$$

- Test statistic

$$F_1 = \frac{MSTR}{MSE}; F_2 = \frac{MSR}{MSE}$$

- provided numerator is bigger than denominator.

Review Self-Practice Problems

- 9.14** Complete the ANOVA table and determine the extent to which this information supports the claim that on an average there are no treatment differences:

Source	df	SS	MSS	F-value
Between	2	—	5	
Within	—	14	—	—
Total	9	—		

- 9.15** A manager obtained the following data on the time (in days) needed to do a job. Use these data to test whether the mean time needed to complete a job differs for four persons. Use $\alpha = 0.05$.

Persons	Time		
1	8	10	9
2	12	16	15
3	15	18	14
4	6	10	7

- 9.16** A leading oil company claims that its engine oil improves engine efficiency. To verify this claim, the company's

brand A is compared with three other competing brands B, C, and D. The data of the survey consists of the km per litre consumption for a combination of city and highway travel, and are as follows:

Size of Container	Brand			
	A	B	C	D
1	36	34	33	35
2	29	26	28	27
3	25	24	25	23
4	19	20	18	18

- (a) Is there any difference in the average mileage for these four brands?

- (b) Is there any difference in the average mileage for a combination of city and highway travel?

- 9.17** A TV manufacturing company claims that the performance of its brand A TV set is better than two other brands. To verify this claim, a sample of 5 TV sets are selected from each brand and the frequency of repair during the first year of purchase is recorded. The results are as under:

TV Brands		
A	B	C
4	7	4
6	4	6
7	3	6
5	6	3
8	5	1

In view of this data, can it be concluded that there is a significant difference between the three brands?

- 9.18 Three varieties of coal were tested for ash content by five different laboratories. The results are as under:

Variety of Coal	Laboratory				
	1	2	3	4	5
A	9	7	6	5	8
B	7	4	5	4	5
C	6	5	6	7	6

In view of this data, can it be concluded that all three varieties of coal have an equal amount of ash content?

- 9.19 An Insurance Company wants to test whether three of its field officers, A, B, and C in a given territory, meet equal number of prospective customers during a given period of time. A record of the previous four months showed the following results for the number of customers contacted by each field officer for each month:

Month	Field Officer		
	A	B	C
1	8	6	14
2	9	8	12
3	11	10	18
4	12	4	8

Is there any significant difference in the average number of contacts made by the three field officers per month?

- 9.20 A departmental store chain is considering opening a new store at one of three locations. An important factor in making such a decision is the household income in these areas. If the average income per household is similar, then the management can choose any one of these three locations. A random survey of various households in each location is undertaken and their annual combined income is recorded. This data is as under:

Annual Household Income (Rs '000s)		
Area 1	Area 2	Area 3
70	100	60
72	110	65
75	108	57
80	112	84
83	113	84
—	120	70
—	100	—

Can the average income per household in these areas be considered to be the same?

- 9.21 Four types of advertising displays were set up in twelve retail outlets, with three outlets randomly assigned to each of the displays. The data on product sales according to the advertising displays are as under:

Types of Display			
A	B	C	D
40	53	48	48
44	54	38	61
43	59	46	47

Does the type of advertising display used at the point of purchase affect the average level of sales?

Hints and Answers

- 9.14 df (within) = total df - df (between) = 9 - 2 = 7, that is, $k - 1 = 2$ and $n - k = 7$

$$MSB = \frac{SSB}{k-1} \quad \text{or} \quad 5 = \frac{SSB}{2}, SSB = 10 \text{ and}$$

$$MSW = \frac{SSW}{n-k} = \frac{14}{7} = 2$$

$$SST = SSB + SSW = 10 + 14 = 24;$$

$$F = \frac{MSB}{MSW} = \frac{5}{2} = 2.5$$

Source	SS	df	MSS	F-value
• Between rows	10	2	5	
• Within column	14	7	2	2.5
Total	24	9		

- 9.15 Let H_0 : (a) No significant difference in efficiency of four persons

(b) Mean time needed to complete the job is equal

Source	SS	df	MSS	F-value
• Between rows	138.667	3	46.22	$F_1 = 47.64$
• Within columns	22.167	2	11.08	$F_2 = 11.42$
• Residual error	5.833	6	0.97	
Total		11		

(a) $F_{cal} = 47.64$ is more than its table value $F_{0.05} = 4.75$ at $df = (3, 6)$, the null hypothesis is rejected.

(b) $F_{cal} = 11.42$ is more than its table value $F_{0.05} = 5.14$ at $df = (2, 6)$, the null hypothesis is rejected.

- 9.16** Let H_0 : (a) No significant difference in average mileage for four brands of engine oils.
 (b) No significant difference in average mileage for a combination of city and highway travel.

Source	SS	df	MSS	F-value
• Between sizes	519.50	3	173.16	$F_1 = \frac{173.16}{1.11} = 156.00$
• Within brands	5.50	3	1.83	$F_2 = \frac{1.83}{1.11} = 1.64$
• Residual error	10.00	9	1.11	
Total	535.00	15		

- (a) $F_{\text{cal}} = 156$ is more than its table value $F_{0.05} = 3.862$, for $df = (3, 9)$, the null hypothesis is rejected.
 (b) $F_{\text{cal}} = 1.64$ is less than its table value $F_{0.05} = 3.862$ for $df = (3, 9)$, the null hypothesis is accepted.

- 9.17** Let H_0 : No significant difference in the performance of three brands of TV sets.

Source	SS	df	MSS	F-value
• Between samples	10	2	5.00	$\frac{5.00}{3.17} = 1.58$
• Within samples	38	12	3.17	
Total	48	14		

Since $F_{\text{cal}} = 1.58$ is less than its table value $F_{0.05} = 3.89$ at $df = (2, 12)$, the null hypothesis is accepted.

- 9.18** Let H_0 : Ash content is equal in all varieties of coal.

Source	SS	df	MSS	F-value
• Between samples	8.673	2	4.337	$\frac{4.337}{1.611} = 2.69$
• Within samples	19.336	12	1.611	
Total	28.01	14		

Since $F_{\text{cal}} = 2.69$ is less than its table value $F_{0.05} = 3.89$ at $df = (2, 12)$, the null hypothesis is accepted.

- 9.19** Let H_0 : All field officers met equal number of customers in the previous four months.

Source	SS	df	MSS	F-value
• Between samples	72	2	36	$\frac{36}{9.1} = 3.95$
• Within samples	10	9	9.1	
Total	82	11		

Since $F_{\text{cal}} = 3.95$ is less than its table value $F_{0.05} = 4.26$ at $df = (2, 9)$, the null hypothesis is accepted.

- 9.20** Let H_0 : No significant difference in the average income per household in all the three areas.

Source	SS	df	MSS	F-value
• Between samples	5787	2	2893.5	$\frac{2893.5}{74.26} = 38.96$
• Within samples	114	15	74.26	
Total	6901	17		

Since $F_{\text{cal}} = 38.96$ is more than its table value $F_{0.05} = 3.68$ at $df = (2, 15)$, the null hypothesis is rejected.

- 9.21** Let H_0 : Type of advertising display used at the point of purchase does not affect the average level of sales.

Source	SS	df	MSS	F-value
• Between types of display	351.5	3	117.2	$\frac{117.2}{25.9} = 4.53$
• Within displays	207.4	8	25.9	
Total	558.9	11		

Since $F_{\text{cal}} = 4.53$ is more than its table value $F_{0.05} = 4.07$ at $df = (3, 8)$, the null hypothesis is rejected.

Case Studies

Case 9.1: FMCG Company

A FMCG company wished to study the effects of four training programmes on the sales abilities of their sales personnel. Thirty-two people were randomly divided into four groups of equal size, and the groups were then subjected to the different sales training programmes. Because there were some dropouts during the training programmes due to illness, vacations, and so on, the number of trainees completing the programmes varied

from group to group. At the end of the training programmes, each salesperson was randomly assigned a sales area from a group of sales areas that were judge to have equivalent sales potentials. The sales made by each of the four groups of salespeople during the first week after completing the training programme are listed in the table:

Training Programme

1	2	3	4
78	99	74	81
84	86	87	63
86	90	80	71
92	93	83	65
69	94	78	86
73	85	73	79
97	70		
91			
<hr/> 482	<hr/> 735	<hr/> 402	<hr/> 588

Questions for Discussion

1. Analyse the experiment using the appropriate method.
2. Identify the treatments or factors of interest to the researcher and investigate any significant effects.
3. What are the practical implications of this experiment?
4. Write a paragraph explaining the results of your analysis.

Nothing is good or bad by comparison.

—Thomas Fuller

Correlation Analysis

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- express quantitatively the degree and direction of the covariation or association between two variables.
- determine the validity and reliability of the covariation or association between two variables.
- provide a test of hypothesis to determine whether a linear relationship actually exists between the variables.

10.1 INTRODUCTION

The statistical methods, discussed so far, are used to analyse the data involving only one variable. Often an analysis of data concerning two or more quantitative variables is needed to look for any statistical relationship or association between them that can describe specific numerical features of the association. The knowledge of such a relationship is important to make inferences from the relationship between variables in a given situation. Few instances where the knowledge of an association or relationship between two variables would prove vital to make decision are:

- Family income and expenditure on luxury items.
- Yield of a crop and quantity of fertilizer used.
- Sales revenue and expenses incurred on advertising.
- Frequency of smoking and lung damage.
- Weight and height of *individuals*.
- Age and sign legibility distance.
- Age and hours of TV viewing per day.

A statistical technique that is used to analyse the strength and direction of the relationship between two quantitative variables, is called *correlation analysis*. A few definitions of correlation analysis are:

- An analysis of the relationship of two or more variables is usually called correlation.
— A. M. Tuttle
- When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation.
— Croxton and Cowden

Coefficient of correlation: A statistical measure of the degree of association between two variables.

The **coefficient of correlation**, is a number that indicates the *strength (magnitude)* and *direction* of statistical relationship between two variables.

- The **strength** of the relationship is determined by the closeness of the points to a straight line when a pair of values of two variables are plotted on a graph. A straight line is used as the frame of reference for evaluating the relationship.
- The **direction** is determined by whether one variable generally increases or decreases when the other variable increases.

The importance of examining the statistical relationship between two or more variables can be divided into the following questions and accordingly requires the statistical methods to answer these questions:

- (i) Is there an association between two or more variables? If yes, what is the form and degree of that relationship?
- (ii) Is the relationship strong or significant enough to be useful to arrive at a desirable conclusion?
- (iii) Can the relationship be used for predictive purposes, that is, to predict the most likely value of a dependent variable corresponding to the given value of independent variable or variables?

In this chapter the first two questions will be answered, while the third question will be answered in Chapter 11.

For correlation analysis, the data on values of two variables must come from sampling in pairs, one for each of the two variables. The pairing relationship should represent some time, place, or condition.

10.2 SIGNIFICANCE OF MEASURING CORRELATION

The objective of any scientific and clinical research is to establish relationships between two or more sets of observations or variables to arrive at some conclusion which is also near to reality. Finding such relationships is often an initial step for identifying causal relationships. Few advantages of measuring an association (or correlation) between two or more variables are as under:

1. Correlation analysis contributes to the understanding of economic behaviour, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilizing forces may become effective. —W. A. Neiswanger
2. The effect of correlation is to reduce the range of uncertainty of our prediction. The prediction based on correlation analysis will be more reliable and near to reality.
— Tippett
3. In economic theory we come across several types of variables which show some kind of relationship. For example, there exists a relationship between price, supply, and quantity demanded; convenience, amenities, and service standards are related to customer retention; yield of a crop related to quantity of fertilizer applied, type of soil, quality of seeds, rainfall, and so on. Correlation analysis helps in quantifying precisely the degree of association and direction of such relationships.
4. Correlations are useful in the areas of healthcare such as determining the validity and reliability of clinical measures or in expressing how health problems are related to certain biological or environmental factors. For example, correlation coefficient can be used to determine the degree of inter-observer reliability for two doctors who are assessing a patient's disease.

10.3 CORRELATION AND CAUSATION

There are at least three criteria for establishing a causal relationship; correlation is one of them. While drawing inferences from the value of correlation coefficient, we overlook the fact that it measures only the strength of a linear relationship and it does not necessarily

imply a causal relationship. That is, there are several other explanations for finding a correlation.

The following factors should be examined to interpret the nature and extent of relationship between two or more variables:

1. **Chance coincidence:** A correlation coefficient may not reach any statistical significance, that is, it may represent a nonsense (spurious) or chance association. For example, (i) a positive correlation between growth in population and wheat production in the country has no statistical significance. Because, each of the two events might have entirely different, unrelated causes. (ii) While estimating the correlation in sales revenue and expenditure on advertisements over a period of time, the investigator must be certain that the outcome is not due to biased sampling or sampling error. That is, he needs to show that a correlation coefficient is statistically significant and not just due to random sampling error.
2. **Influence of third variable:** If the correlation coefficient does not establish any relationship, it can be used as a source for testing null and alternative hypotheses about a population. For example, it has been proved that smoking causes lung damage. However, given that there is often multiple reasons of health problems, the reason of stress cannot be ruled out. Similarly, there is a positive correlation between the yield of rice and tea because the crops are influenced by the amount of rainfall. But the yield of any one is not influenced by other.
3. **Mutual influence:** There may be a high degree of relationship between two variables but it is difficult to say as to which variable is influencing the other. For example, variables like price, supply, and demand of a commodity are mutually correlated. According to the principle of economics, as the price of a commodity increases, its demand decreases, so price influences the demand level. But if demand of a commodity increases due to growth in population, then its price also increases. In this case increased demand make an effect on the price. However, the amount of export of a commodity is influenced by an increase or decrease in custom duties but the reverse is normally not true.

10.4 TYPES OF CORRELATIONS

There are three broad types of correlations:

1. Positive and negative,
2. Linear and non-linear,
3. Simple, partial, and multiple.

In this chapter we will discuss simple linear positive or negative correlation analysis.

10.4.1 Positive and Negative Correlation

A positive (or direct) correlation refers to the same direction of change in the values of variables. In other words, if values of variables are varying (i.e., increasing or decreasing) in the same direction, then such correlation is referred to as **positive correlation**.

A **negative (or inverse) correlation** refers to the change in the values of variables in opposite direction.

The following examples illustrate the concept of positive and negative correlation.

Positive Correlation

Increasing → x :	5	8	10	15	17
Increasing → y :	10	12	16	18	20
Decreasing → x :	17	15	10	8	5
Decreasing → y :	20	18	16	12	10

Negative Correlation

Increasing → x :	5	8	10	15	17
Decreasing → y :	20	18	16	12	10
Decreasing → x :	17	15	12	10	6
Increasing → y :	2	7	9	13	14

It may be noted here that the change (increasing or decreasing) in values of both the variables is not proportional or fixed.

10.4.2 Linear and Non-Linear Correlation

A linear correlation implies a constant change in one of the variable values with respect to a change in the corresponding values of another variable. In other words, a correlation is referred to as *linear correlation* when variations in the values of two variables have a constant ratio. The following example illustrates a linear correlation between two variables x and y .

x :	10	20	30	40	50
y :	40	60	80	100	120

When these pairs of values of x and y are plotted on a graph paper, the line joining these points would be a straight line.

A non-linear (or curvi-linear) correlation implies an absolute change in one of the variable values with respect to changes in values of another variable. In other words, a correlation is referred to as a *non-linear correlation* when the amount of change in the values of one variable does not bear a constant ratio to the amount of change in the corresponding values of another variable. The following example illustrates a non-linear correlation between two variables x and y .

x :	8	9	9	10	10	28	29	30
y :	80	130	170	150	230	560	460	600

When these pair of values of x and y are plotted on a graph paper, the line joining these points would not be a straight line, rather it would be curvi-linear.

10.4.3 Simple, Partial, and Multiple Correlation

The distinction between simple, partial, and multiple correlation is based upon the number of variables involved in the correlation analysis.

If only two variables are chosen to study correlation between them, then such a correlation is referred to as *simple correlation*. A study on the yield of a crop with respect to only amount of fertilizer, or sales revenue with respect to amount of money spent on advertisement, are a few examples of simple correlation.

In *partial correlation*, two variables are chosen to study the correlation between them, but the effect of other influencing variables is kept constant. For example (i) yield of a crop is influenced by the amount of fertilizer applied, rainfall, quality of seed, type of soil, and pesticides, (ii) sales revenue from a product is influenced by the level of advertising expenditure, quality of the product, price, competitors, distribution, and so on. In such cases an attempt to measure the correlation between yield and seed quality, assuming that the average values of other factors exist, becomes a problem of partial correlation.

In *multiple correlation*, the relationship between more than three variables is considered simultaneously for study. For example, employer-employee relationship in any organization may be examined with reference to, training and development facilities; medical, housing, and education to children facilities; salary structure; grievances handling system; and so on.

10.5 METHODS OF CORRELATION ANALYSIS

The correlation between two ratio-scaled (numeric) variables is represented by the letter r which takes on values between -1 and $+1$ only. Sometimes this measure is called the '**Pearson product moment correction**' or the **correlation coefficient**. The correlation coefficient is scale free and therefore its interpretation is independent of the units of measurement of two variables, say x and y .

In this chapter, the following methods of finding the correlation coefficient between two variables x and y are discussed:

1. Scatter Diagram method
2. Karl Pearson's Coefficient of Correlation method
3. Spearman's Rank Correlation method
4. Method of Least-squares

Figure 10.1 shows how the strength of the association between two variables is represented by the coefficient of correlation.

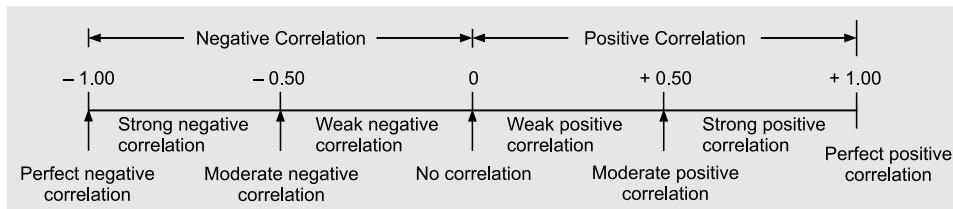


Figure 10.1
Interpretation of Correlation Coefficient

10.5.1 Scatter Diagram Method

The **scatter diagram** method is a quick at-a-glance method of determining of an apparent relationship between two variables, if any. A scatter diagram (or a graph) can be obtained on a graph paper by plotting observed (or known) pairs of values of variables x and y , taking the independent variable values on the x -axis and the dependent variable values on the y -axis.

It is common to try to draw a straight line through data points so that an equal number of points lie on either side of the line. The relationship between two variables x and y described by the data points is defined by this straight line.

In a scatter diagram the horizontal and vertical axes are scaled in units corresponding to the variables x and y , respectively. Figure 10.2 shows examples of different types of relationships based on pairs of values of x and y in a sample data. The pattern of data points in the diagram indicates that the variables are related. If the variables are related, then the dotted line appearing in each diagram describes relationship between the two variables.

The patterns depicted in Fig. 10.2(a) and (b) represent linear relationships since the patterns are described by straight lines. The pattern in Fig. 10.2(a) shows a *positive* relationship since the value of y tends to increase as the value of x increases, whereas pattern in Fig. 10.2(b) shows a *negative* relationship since the value of y tends to decrease as the value of x increases.

The pattern depicted in Fig. 10.2(c) illustrates very low or no relationship between the values of x and y , whereas Fig. 10.2(d) represents a curvilinear relationship since it is described by a curve rather than a straight line. Figure 10.2(e) illustrates a positive linear relationship with a widely scattered pattern of points. The wider scattering indicates that there is a lower degree of association between the two variables x and y than there is in Fig. 10.2(a).

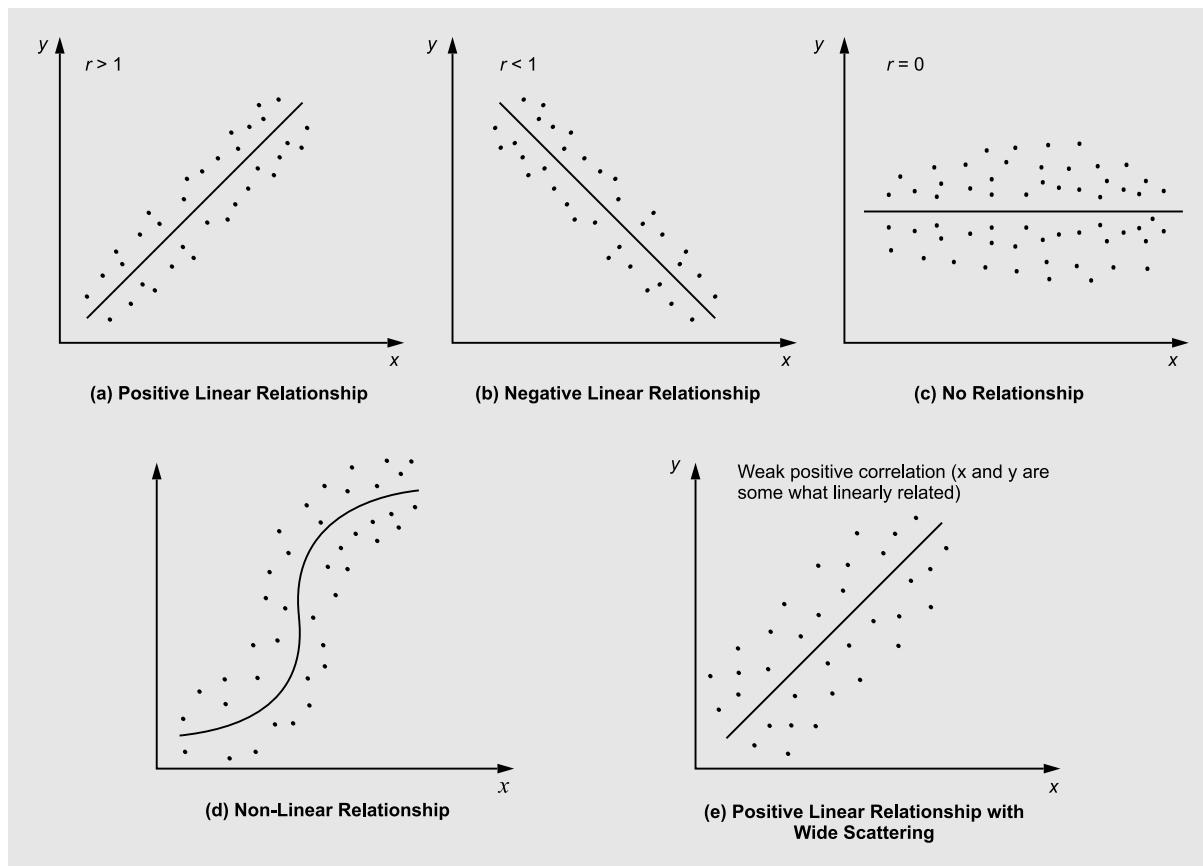
Scatter diagram: A graph of pairs of values of two variables that is plotted to indicate a visual display of the pattern of their relationship.

Interpretation of Correlation Coefficients While interpreting correlation coefficient r , the following points should be taken into account:

- (i) A low value of r does not indicate that the variables are unrelated but indicates that the relationship is poorly described by a straight line. A non-linear relationship may also exist.
- (ii) A correlation does not imply a *cause-and-effect* relationship, it is merely an observed association.

Figure 10.2

Typical Examples of Correlation Coefficient



Types of Correlation Coefficients Table 10.1 shows several types of correlation coefficients used in statistics along with the conditions of their use. All of them are appropriate for quantifying linear relationship between two variables x and y .

Table 10.1: Types of Correlation Coefficients

Coefficient	Conditions Applied for Use
• ϕ (phi)	Both x and y variables are measured on a nominal scale
• ρ (rho)	Both x and y variables are measured on, or changed to, ordinal scales (rank data)
• r	Both x and y variables are measured on an interval or ratio scale (numeric data)

The correlation coefficient, denoted by η (eta) is used for quantifying nonlinear relationships (It is beyond the scope of this text). In this chapter we will calculate only the commonly used Pearson's r and Spearman's ρ correlation coefficients.

Specific Features of the Correlation Coefficient Regardless of the type of correlation coefficient we use, the following are the common among all of them.

- (i) The value of r depends on the slope of the line passing through the data points and the scattering of the pair of values of variables x and y about this line (for detail see Chapter 11).
- (ii) The sign of the correlation coefficient indicates the direction of the relationship. A positive correlation denoted by + (positive sign) indicates that the two variables tend to increase (or decrease) together (a positive association) and a negative correlation by - (minus sign) indicates that when one variable increases the other is likely to decrease (a negative association).
- (iii) The values of the correlation coefficient range from +1 to -1 regardless of the units of measurements of x and y .
- (iv) The value of $r = +1$ or -1 indicates that there is a perfect linear relationship between two variables, x and y . A perfect correlation implies that every observed pair of values of x and y falls on the straight line.
- (v) The magnitude of the correlation indicates the strength of the relationship, which is the overall closeness of the points to a straight line. The sign of the correlation does indicate about the strength of the linear relationship.
- (vi) Correlation coefficient is independent of the change of origin and scale of reference. In other words, its value remains unchanged when we subtract some constant from every given value of variables x and y (change of origin) and when we divide or multiply by some constant every given value of x and y (change of scale).
- (vii) Correlation coefficient is a pure number independent of the unit of measurement.
- (viii) The value of $r = 0$ indicates that the straight line through the data is exactly horizontal, so that the value of variable x does not change the predicated value of variable y .
- (ix) The square of r , i.e., r^2 is referred to as *coefficient of determination*.

Further, from Fig. 10.2(a) to (e) we conclude that the closer the value of r is to either +1 or -1, the stronger is the association between x and y . Also, closer the value of r to 0, the weaker the association between x and y appears to be.

Example 10.1: Given the following data:

Student	:	1	2	3	4	5	6	7	8	9	10
Management aptitude score	:	400	675	475	350	425	600	550	325	675	450
Grade point average	:	1.8	3.8	2.8	1.7	2.8	3.1	2.6	1.9	3.2	2.3

- (a) Draw this data on a graph paper.
- (b) Is there any correlation between per capita national income and per capita consumer expenditure? If yes, what is your opinion.

Solution: By taking an appropriate scale on the x and y axes, the pair of observations are plotted on a graph paper as shown in Fig. 10.3. The scatter diagram in Fig. 10.3 with straight line represents the relationship between x and y 'fitted' through it.

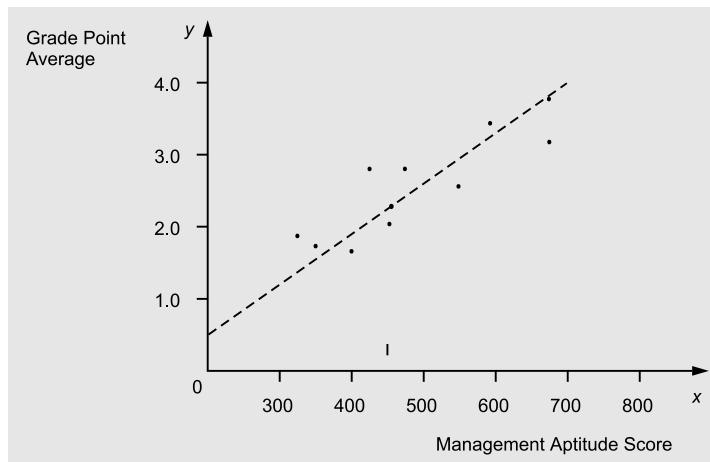


Figure 10.3
Scatter Diagram

Interpretation: From the scatter diagram shown in Fig. 10.3, it appears that there is a high degree of association between two variable values. It is because the data points are very close to a straight line passing through the points. This pattern of dotted points also indicates a high degree of linear positive correlation.

10.5.2 Karl Pearson's Correlation Coefficient

Karl Pearson's correlation coefficient measures quantitatively the extent to which two variables x and y are correlated. For a set of n pairs of values of x and y , Pearson's correlation coefficient r is given by

$$r = \frac{\text{Covariance } (x, y)}{\sqrt{\text{var } x} \sqrt{\text{var } y}} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

where

$$\text{Cov}(x, y) = \frac{1}{n} \sum (x - \bar{x})(y - \bar{y})$$

$$\sigma_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} \quad \leftarrow \text{standard deviation of sample data on variable } x$$

$$\sigma_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n}} \quad \leftarrow \text{standard deviation of sample data on variable } y$$

Substituting mathematical formula for $\text{Cov}(x, y)$ and σ_x and σ_y , we have

$$r = \frac{\frac{1}{n} \sum (x - \bar{x})(y - \bar{y})}{\sqrt{\frac{\sum (x - \bar{x})^2}{n}} \sqrt{\frac{\sum (y - \bar{y})^2}{n}}} = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}} \quad (10-1)$$

Step Deviation Method for Ungrouped Data When actual mean values \bar{x} and \bar{y} are in fraction, the calculation of Pearson's correlation coefficient can be simplified by taking deviations of x and y values from their assumed means A and B , respectively. That is, $d_x = x - A$ and $d_y = y - B$, where A and B are assumed means of x and y values. The formula (10-1) becomes

$$r = \frac{n \sum d_x d_y - (\sum d_x)(\sum d_y)}{\sqrt{n \sum d_x^2 - (\sum d_x)^2} \sqrt{n \sum d_y^2 - (\sum d_y)^2}} \quad (10-2)$$

Step Deviation Method for Grouped Data When data on x and y values are classified or grouped into a frequency distribution, the formula (10-2) is modified as:

$$r = \frac{n \sum f d_x d_y - (\sum f d_x)(\sum f d_y)}{\sqrt{n \sum f d_x^2 - (\sum f d_x)^2} \sqrt{n \sum f d_y^2 - (\sum f d_y)^2}} \quad (10-3)$$

Assumptions of Using Pearson's Correlation Coefficient

- (i) Pearson's correlation coefficient is appropriate to calculate when both variables x and y are measured on an interval or a ratio scale.
- (ii) Both variables x and y are normally distributed, and that there is a linear relationship between these variables.
- (iii) The correlation coefficient is largely affected due to truncation of the range of values in one or both of the variables. This occurs when the distributions of both the variables greatly deviate from the normal shape.
- (iv) There is a cause and effect relationship between two variables that influences the distributions of both the variables. Otherwise correlation coefficient might either be extremely low or even zero.

Advantage and Disadvantages of Pearson's Correlation Coefficient The correlation coefficient is a numerical number between -1 and 1 that summarizes the magnitude as well

as direction (positive or negative) of association between two variables. The chief limitations of Pearson's method are:

- (i) The correlation coefficient always assumes a linear relationship between two variables, whether it is true or not.
- (ii) Great care must be exercised in interpreting the value of this coefficient as very often its value is misinterpreted.
- (iii) The value of the coefficient is unduly affected by the extreme values of two variable values.
- (iv) As compared with other methods the computational time required to calculate the value of r using Pearson's method is lengthy.

10.5.3 Probable Error and Standard Error of Coefficient of Correlation

The probable error (PE) of coefficient of correlation indicates extent to which its value depends on the condition of random sampling. If r is the calculated value of correlation coefficient in a sample of n pairs of observations, then the standard error SE_r of the correlation coefficient r is given by

$$SE_r = \frac{1-r^2}{\sqrt{n}}$$

The probable error of the coefficient of correlation is calculated by the expression:

$$PE_r = 0.6745 SE_r = 0.6745 \frac{1-r^2}{\sqrt{n}}$$

Thus with the help of PE_r we can determine the range within which population coefficient of correlation is expected to fall using following formula:

$$\rho = r \pm PE_r$$

where ρ (rho) represents population coefficient of correlation.

Remarks

1. If $r < PE_r$ then the value of r is not significant, that is, there is no relationship between two variables of interest.
2. If $r > 6PE_r$ then value of r is significant, that is, there exists a relationship between two variables.

Illustration: If $r = 0.8$ and $n = 25$, then PE_r is

$$PE_r = 0.6745 \frac{1-(0.8)^2}{\sqrt{25}} = 0.6745 \frac{0.36}{5} = 0.048$$

Thus the limits within which population correlation coefficient (ρ_r) should fall are

$$r \pm PE_r = 0.8 \pm 0.048 \quad \text{or} \quad 0.752 \leq \rho_r \leq 0.848$$

10.5.4 The Coefficient of Determination

The squared value of the correlation coefficient r is called **coefficient of determination**, denoted as r^2 . It always has a value between 0 and 1. By squaring the correlation coefficient we retain information about the strength of the relationship but we lose information about the direction. This measure represents the proportion (or percentage) of the total variability of the dependent variable, y that is accounted for or explained by the independent variable, x . The proportion (or percentage) of variation in y that x can explain determines more precisely the extent or strength of association between two variables x and y (see Chapter 11 for details).

- The coefficient of correlation r has been grossly overrated and is used entirely too much. Its square, coefficient of determination r^2 , is a much more useful measure of the linear covariation of two variables. The reader should develop the habit of squaring every correlation coefficient he finds cited or stated before coming to any conclusion about the extent of the linear relationship between two correlated variables. —Tuttle

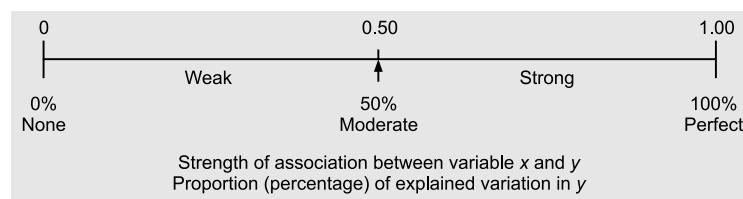
Coefficient of determination:

A statistical measure of the proportion of the variation in the dependent variable that is explained by independent variable.

Interpretation of Coefficient of Determination: Coefficient of determination is preferred for interpreting the strength of association between two variables because it is easier to interpret a percentage. Figure 10.4 illustrates the meaning of the coefficient of determination:

- If $r^2 = 0$, then *no variation* in y can be *explained* by the variable x . It is shown in Fig 10.2(c) where x is of no value in predicting the value of y . There is *no association* between x and y .
- If $r^2 = 1$, then values of y are *completely explained* by x . There is *perfect association* between x and y .
- If $0 \leq r^2 \leq 1$, the degree of explained variation in y as a result of *variation in values of x* depends on the value of r^2 . Value of r^2 closer to 0 shows low proportion of variation in y explained by x . On the other hand value of r^2 closer to 1 show that variable x can predict the actual value of the variable y .

Figure 10.4
Interpretation of Coefficient of Determination



Mathematically, the coefficient of determination is given by

$$\begin{aligned} r^2 &= 1 - \frac{\text{Explained variability in } y}{\text{Total variability in } y} \\ &= 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} = 1 - \frac{n \sum y^2 - a \sum y - b \sum xy}{n \sum y^2 - (\bar{y})^2} \end{aligned}$$

where $\hat{y} = a + bx$ is the estimated value of y for given values of x . One minus the ratio between these two variations is referred as the *coefficient of determination*.

For example, let correlation between variable x (height) and variable y (weight) be $r = 0.70$. Now the coefficient of determination $r^2 = 0.49$ or 49 per cent, implies that only 49 per cent of the variation in variable y (weight) can be accounted for in terms of variable x (height). The remaining 51 per cent of the variability may be due to other factors, say for instance, tendency to eat fatty foods.

It may be noted that even a relatively high correlation coefficient $r = 0.70$ accounts for less than 50 per cent of the variability. In this context, it is important to know that 'variability' refers to how values of variable y are scattered around its own mean value. That is, as in the above example, some people will be heavy, some average, some light. So we can account for 49 per cent of the total variability of weight (y) in terms of height (x) if $r=0.70$. The greater the correlation coefficient, the greater the coefficient of determination, and the variability in dependent variable can be accounted for in terms of independent variable.

Example 10.2: The following table gives indices of industrial production and number of registered unemployed people (in lakh). Calculate the value of the correlation coefficient.

Year	:	1991	1992	1993	1994	1995	1996	1997	1998
Index of Production	:	100	102	104	107	105	112	103	99
Number Unemployed	:	15	12	13	11	12	12	19	26

Solution: Calculations of Karl Pearson's correlation coefficient are shown in the table below:

Year	Production	$dx = (x - \bar{x})$	d_x^2	Unemployed	$d_y^2 = (y - \bar{y})$	d_y^2	$d_x d_y$
	x			y			
1991	100	-4	16	15	0	0	0
1992	102	-2	4	12	-3	9	+6
1993	104	0	0	13	-2	4	0
1994	107	+3	9	11	-4	16	-12
1995	105	+1	1	12	-3	9	-3
1996	112	+8	64	12	-3	9	-24
1997	103	-1	1	19	+4	16	-4
1998	99	-5	25	26	+11	121	-55
Total	832	0	120	120	0	184	-92

$$\bar{x} = \frac{\sum x}{n} = \frac{832}{8} = 104; \quad \bar{y} = \frac{\sum y}{n} = \frac{120}{8} = 15$$

Applying the formula, $r = \frac{n \sum d_x d_y - (\sum d_x)(\sum d_y)}{\sqrt{n \sum d_x^2 - (\sum d_x)^2} \sqrt{n \sum d_y^2 - (\sum d_y)^2}}$

$$= \frac{8 \times -92}{\sqrt{8 \times 120} \sqrt{8 \times 184}} = \frac{-92}{10.954 \times 13.564}$$

$$= \frac{-92}{148.580} = -0.619$$

Interpretation: Since coefficient of correlation $r = -0.619$ is moderately negative, it indicates that there is a moderately large inverse correlation between the two variables. Hence we conclude that as the production index increases, the number of unemployed decreases and vice-versa.

Example 10.3: The following table gives the distribution of items of production and also the relatively defective items among them, according to size groups. Find the correlation coefficient between size and defect in quality.

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 defective items	:	150	162	170	180	180	114

[Delhi Univ., BCom, 1999]

Solution: Let group size be denoted by variable x and number of defective items by variable y . Calculations for Karl Pearson's correlation coefficient are shown below:

Size-Group	Mid-value m	$d_x = m - 17.5$	d_x^2	Percent of Defective Items	$d_y = y - 50$	d_y^2	$d_x d_y$
15–16	15.5	-2	4	75	+25	625	-50
16–17	16.5	-1	1	60	+10	100	-10
17–18	17.5	0	0	50	0	0	0
18–19	18.5	+1	1	50	0	0	0
19–20	19.5	+2	4	45	-5	25	-10
20–21	20.5	+3	9	38	-12	144	-36
		3	19		18	894	-106

Substituting values in the formula of Karl Pearson's correlation coefficient r , we have

$$r = \frac{n \sum d_x d_y - (\sum d_x)(\sum d_y)}{\sqrt{n \sum d_x^2 - (\sum d_x)^2} \sqrt{n \sum d_y^2 - (\sum d_y)^2}}$$

$$\begin{aligned}
 &= \frac{6 \times -106 - 3 \times 18}{\sqrt{6 \times 19 - (3)^2} \sqrt{6 \times 894 - (18)^2}} = \frac{-636 - 54}{\sqrt{105} \sqrt{5040}} \\
 &= -\frac{690}{727.46} = -0.949
 \end{aligned}$$

Interpretation: Since value of r is negative, and is moderately close to -1 , statistical association between x (size group) and y (percent of defective items) is moderate and negative, we conclude that when size of group increases, the number of defective items decreases and vice-versa.

Example 10.4: The following data relate to age of employees and the number of days they reported sick in a month.

Employees :	1	2	3	4	5	6	7	8	9	10
Age :	30	32	35	40	48	50	52	55	57	61
Sick days :	1	0	2	5	2	4	6	5	7	8

Calculate Karl Pearson's coefficient of correlation and interpret it.

[Kashmir Univ., BCom, 1997]

Solution: Let age and sick days be represented by variables x and y , respectively. Calculations for value of correlation coefficient are shown below:

<i>Age</i>	<i>Sick days</i>					
	x	$dx = x - \bar{x}$	d_x^2	y	$d_y = y - \bar{y}$	d_y^2
30	-16	256	1	-3	9	48
32	-14	196	0	-4	16	56
35	-11	121	2	-2	4	22
40	-6	36	5	1	1	-6
48	2	4	2	-2	4	-4
50	4	16	4	0	0	0
52	6	36	6	2	4	12
55	9	81	5	1	1	9
57	11	121	7	3	9	33
61	15	225	8	4	16	60
<hr/>	<hr/>	<hr/>	<hr/>	<hr/>	<hr/>	<hr/>
460	0	1092	40	0	64	230

$$\bar{x} = \frac{\sum x}{n} = \frac{460}{10} = 46 \text{ and } \bar{y} = \frac{\sum y}{n} = \frac{40}{10} = 4$$

Substituting values in the formula of Karl Pearson's correlation coefficient r , we have

$$\begin{aligned}
 r &= \frac{n \sum d_x d_y - (\sum d_x)(\sum d_y)}{\sqrt{n \sum d_x^2 - (\sum d_x)^2} \sqrt{n \sum d_y^2 - (\sum d_y)^2}} = \frac{10 \times 230}{\sqrt{10 \times 1092} \sqrt{10 \times 64}} \\
 &= \frac{230}{264.363} = 0.870
 \end{aligned}$$

Interpretation: Since value of r is positive, therefore age of employees and number of sick days are positively correlated to a high degree. Hence we conclude that as the age of an employee increases, he is likely to go on sick leave more often than others.

Example 10.5: The following table gives the frequency, according to the marks, obtained by 67 students in an intelligence test. Measure the degree of relationship between age and marks:

Test Marks	Age in years				Total
	18	19	20	21	
200–250	4	4	2	1	11
250–300	3	5	4	2	14
300–350	2	6	8	5	21
350–400	1	4	6	10	21
Total	10	19	20	18	67

[Allahabad Univ., BCom, 1999]

Solution: Let age of students and marks obtained by them be represented by variables x and y , respectively. Calculations for correlation coefficient for this bivariate data is shown below:

		Age in years				Total, f	fd_y	fd_y^2	$fd_x d_y$
x	d_x	18	19	20	21				
200–250	-1	(4)	(0)	(-2)	(-2)	11	-11	11	0
	4	4	2	1	1				
250–300	0	(0)	(0)	(0)	(0)	14	0	0	0
	3	5	4	2					
300–350	1	(-2)	(0)	(8)	(10)	21	21	21	16
	2	6	8	5					
350–400	2	(-2)	(0)	(12)	(40)	21	42	84	50
	1	4	6	10					
Total, f		10	19	20	18	$n = 67$	$\Sigma fd_y = 52$	$\Sigma fd_y^2 = 116$	$\Sigma fd_x d_y = 66$
fd_x	-10	0	20	36	$\Sigma fd_x = 46$				
fd_x^2	10	0	20	72	$\Sigma fd_x^2 = 102$				
$fd_x d_y$	0	0	18	48	$\Sigma fd_x d_y = 66$				

where $d_x = x - 19$, $d_y = (m - 275)/50$

Substituting values in the formula of Karl Pearson's correlation coefficient, we have

$$\begin{aligned}
 r &= \frac{n \Sigma fd_x d_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{n \Sigma fd_x^2 - (\Sigma fd_x)^2} \sqrt{n \Sigma fd_y^2 - (\Sigma fd_y)^2}} = \frac{67 \times 66 - 46 \times 52}{\sqrt{67 \times 102 - (46)^2} \sqrt{67 \times 116 - (52)^2}} \\
 &= \frac{4422 - 2392}{\sqrt{6834 - 2116} \sqrt{7772 - 2704}} = \frac{2030}{\sqrt{4718} \sqrt{5068}} \\
 &= \frac{2030}{68.688 \times 71.19} = 0.415
 \end{aligned}$$

Interpretation: Since the value of r is positive, therefore age of students and marks obtained in an intelligence test are positively correlated to the extent of 0.415. Hence, we conclude that as the age of students increases, score of marks in intelligence test also increases.

Example 10.6: Calculate the coefficient of correlation from the following bivariate frequency distribution:

Sales Revenue (Rs in lakh)	Advertising Expenditure (Rs in '000)				Total
	5-10	10-15	15-20	20-25	
75-125	4	1	—	—	5
125-175	7	6	2	1	16
175-225	1	3	4	2	10
225-275	1	1	3	4	9
Total	13	11	9	7	40

[Delhi Univ., MBA, 1997]

Solution: Let advertising expenditure and sales revenue be represented by variables x and y , respectively. The calculations for correlation coefficient are shown below:

Revenue y	Mid-value (m) d_x	$x \rightarrow$ Mid-value (m) d_y	Advertising Expenditure				Total, f	fd_y	fd_y^2	$fd_x d_y$
			5-10	10-15	15-20	20-25				
75-125	100	-2	(8) 4	(0) 1	(0) —	(0) —	5	-10	20	8
125-175	150	-1	(7) 7	(0) 6	(-2) 2	(-2) 1	16	-16	16	3
175-225	200	0	(0) 1	(0) 3	(0) 4	(0) 2	10	0	0	0
225-275	250	1	(-1) 1	(0) 1	(3) 3	(8) 4	9	9	9	10
Total, f			13	11	9	7	n = 40	$\Sigma d_y = -17$	$\Sigma d_y^2 = 45$	$\Sigma fd_x d_y = 21$
fd_x			-13	0	9	14	$\Sigma fd_x = 10$			
fd_x^2			13	0	9	28	$\Sigma fd_x^2 = 50$			
$fd_x d_y$			14	0	1	6	$\Sigma fd_x d_y = 21$			

where, $d_x = (m - 12.5)/5$ and $d_y = (m - 200)/50$

Substituting values in the formula of Karl Pearson's correlation coefficient, we have

$$\begin{aligned}
 r &= \frac{n \Sigma fd_x d_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{n \Sigma fd_x^2 - (\Sigma fd_x)^2} \sqrt{n \Sigma fd_y^2 - (\Sigma fd_y)^2}} = \frac{40 \times 21 - 10 \times -17}{\sqrt{40 \times 50 - (10)^2} \sqrt{40 \times 45 - (-17)^2}} \\
 &= \frac{840 + 170}{\sqrt{1900} \sqrt{1511}} = \frac{1010}{1694.373} = 0.596
 \end{aligned}$$

Interpretation: Since the value of r is positive, advertising expenditure and sales revenue are positively correlated to the extent of 0.596. Hence we conclude that as expenditure on advertising increases, the sales revenue also increases.

Example 10.7: A computer, while calculating the correlation coefficient between two variables x and y from 25 pairs of observations, obtained the following results:

$$n = 25, \Sigma x = 125, \Sigma x^2 = 650 \text{ and } \Sigma y = 100, \Sigma y^2 = 460, \Sigma xy = 508$$

It was, however, discovered at the time of checking that he had copied down two pairs of observations as:

x	y		x	y
6	14	instead of	8	12
8	6		6	8

Obtain the correct value of correlation coefficient between x and y .

[MD Univ., MCom, 1998; Kumaon Univ., MBA, 2000]

Solution: The corrected values for terms needed in the formula of Pearson's correlation coefficient are determined as follows:

$$\text{Correct } \Sigma x = 125 - (6 + 8 - 8 - 6) = 125$$

$$\text{Correct } \Sigma y = 100 - (14 + 6 - 12 - 8) = 100$$

$$\begin{aligned} \text{Correct } \Sigma x^2 &= 650 - \{(6)^2 + (8)^2 - (8)^2 - (6)^2\} \\ &= 650 - \{36 + 64 - 64 - 36\} = 650 \end{aligned}$$

$$\begin{aligned} \text{Correct } \Sigma y^2 &= 460 - \{(14)^2 + (6)^2 - (12)^2 - (8)^2\} \\ &= 460 - \{196 + 36 - 144 - 64\} = 436 \end{aligned}$$

$$\begin{aligned} \text{Correct } \Sigma xy &= 508 - \{(6 \times 14) + (8 \times 6) - (8 \times 12) - (6 \times 8)\} \\ &= 508 - \{84 - 48 - 96 - 48\} = 520 \end{aligned}$$

Applying the formula

$$\begin{aligned} r &= \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n \Sigma x^2 - (\Sigma x)^2} \sqrt{n \Sigma y^2 - (\Sigma y)^2}} = \frac{25 \times 520 - 125 \times 100}{\sqrt{25 \times 650 - (125)^2} \sqrt{25 \times 436 - (100)^2}} \\ &= \frac{13,000 - 12,500}{\sqrt{625} \sqrt{900}} = \frac{500}{25 \times 30} = 0.667 \end{aligned}$$

Thus, the correct value of correlation coefficient between x and y is 0.667.

Self-Practice Problems 10A

- 10.1** Making use of the data summarized below, calculate the coefficient of correlation.

Case	x_1	x_2	Case	x_1	x_2
A	10	9	E	12	11
B	6	4	F	13	13
C	9	6	G	11	8
D	10	9	H	9	4

- 10.2** Find the correlation coefficient by Karl Pearson's method between x and y and interpret its value.

$$\begin{aligned} x : 57 & 42 & 40 & 33 & 42 & 45 & 42 & 44 & 40 & 56 & 44 & 43 \\ y : 10 & 60 & 30 & 41 & 29 & 27 & 27 & 19 & 18 & 19 & 31 & 29 \end{aligned}$$

- 10.3** Calculate the coefficient of correlation from the following data:

$$\begin{aligned} x : 100 & 200 & 300 & 400 & 500 & 600 & 700 \\ y : 30 & 50 & 60 & 80 & 100 & 110 & 130 \end{aligned}$$

- 10.4** Calculate the coefficient of correlation between x and y from the following data and calculate the probable errors. Assume 69 and 112 as the mean value for x and y respectively.

$$\begin{aligned} x : 78 & 89 & 99 & 60 & 50 & 79 & 68 & 61 \\ y : 125 & 137 & 156 & 112 & 107 & 136 & 123 & 108 \end{aligned}$$

- 10.5** Find the coefficient of correlation from the following data:

$$\begin{aligned} \text{Cost : } & 39 & 65 & 62 & 90 & 82 & 75 & 25 & 98 & 36 & 78 \\ \text{Sales : } & 47 & 53 & 58 & 86 & 62 & 68 & 60 & 91 & 51 & 84 \end{aligned}$$

[Madras Univ., BCom, 1997]

- 10.6** Calculate Karl Pearson's coefficient of correlation between age and playing habits from the data given below. Also calculate the probable error and comment on the value:

$$\begin{aligned} \text{Age : } & 20 & 21 & 22 & 23 & 24 & 25 \\ \text{No. of students : } & 500 & 400 & 300 & 240 & 200 & 160 \\ \text{Regular players : } & 400 & 300 & 180 & 96 & 60 & 24 \end{aligned}$$

[HP Univ., MBA, 1997]

- 10.7** Find the coefficient of correlation between age and the sum assured (in 1000 Rs) from the following table:

Age Group (years)	Sum Assured (in Rs)				
	10	20	30	40	50
20-30	4	6	3	7	1
30-40	2	8	15	7	1
40-50	3	9	12	6	2
50-60	8	4	2	—	—

[Delhi Univ., MBA, 1999]

- 10.8** Family income and its percentage spent on food in the case of one hundred families gave the following bivariate frequency distribution. Calculate the coefficient of correlation and interpret its value.

Food Expenditure (in percent)	Monthly Family Income (Rs)				
	2000– 3000	3000– 4000	4000– 5000	5000– 6000	6000– 7000
10–15	—	—	—	3	7
15–20	—	4	9	4	3
20–25	7	6	12	5	—
25–30	3	10	19	8	—

[Delhi Univ., MBA, 2000]

- 10.9** With the following data in 6 cities, calculate Pearson's

coefficient of correlation between the density of population and death rate:

City	Area in Kilometres	Population (in '000)	No. of Deaths
A	150	30	300
B	180	90	1440
C	100	40	560
D	60	42	840
E	120	72	1224
F	80	24	312

[Sukhadia Univ., BCom, 1998]

- 10.10** The coefficient of correlation between two variables x and y is 0.3. The covariance is 9. The variance of x is 16. Find the standard deviation of y series.

Hints and Answers

10.1 $\bar{x}_1 = 80/8 = 10$, $\bar{x}_2 = 64/8 = 8$;

$$r = \frac{43}{\sqrt{32} \sqrt{72}} = 0.896$$

10.2 $r = -0.554$ **10.3** $r = 0.997$

10.4 $r = 0.014$ **10.5** $r = 0.780$

10.6 $r = 0.005$

10.8 $r = -0.438$

10.7 $r = -0.256$

10.9 $r = 0.988$

10.10 Given $\sigma_x = \sqrt{16} = 4$; $r = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$

or $0.3 = \frac{9}{4\sigma_y}$ or $\sigma_y = 7.5$.

10.5.5 Spearman's Rank Correlation Coefficient

This method of finding the correlation coefficient between two variables was developed by the British psychologist Charles Edward Spearman in 1904. This method is applied to measure the association between two variables when only *ordinal (or rank) data* are available. In other words, this method is applied in a situation in which quantitative measure of certain qualitative factors such as judgement, brands personalities, TV programmes, leadership, colour, taste, cannot be fixed, but individual observations can be arranged in a definite order (also called rank). The ranking is decided by using a set of ordinal rank numbers, with 1 for the individual observation ranked first either in terms of quantity or quality; and n for the individual observation ranked last in a group of n pairs of observations. Mathematically, Spearman's rank correlation coefficient is defined as:

$$R = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \quad (10-4)$$

where R = rank correlation coefficient

R_1 = rank of observations with respect to first variable

R_2 = rank of observations with respect to second variable

$d = R_1 - R_2$, difference in a pair of ranks

n = number of paired observations or individuals being ranked

The number '6' is placed in the formula as a scaling device, it ensures that the possible range of R is from -1 to 1. While using this method we may come across three types of cases.

Advantages and Disadvantages of Spearman's Correlation Coefficient Method

Advantages

- (i) This method is easy to understand and its application is simpler than Pearson's method.
- (ii) This method is useful for correlation analysis when variables are expressed in qualitative terms like beauty, intelligence, honesty, efficiency, and so on.
- (iii) This method is appropriate to measure the association between two variables if the data type is at least ordinal scaled (ranked)
- (iv) The sample data of values of two variables is converted into ranks either in ascending order or descending order for calculating degree of correlation between two variables.

Disadvantages

- (i) Values of both variables are assumed to be normally distributed and describing a linear relationship rather than non-linear relationship.
- (ii) A large computational time is required when number of pairs of values of two variables exceed 30.
- (iii) This method cannot be applied to measure the association between two variable grouped data.

Case I: When Ranks are Given

When observations in a data set are already arranged in a particular order (rank), take the differences in pairs of observations to determine d . Square these differences and obtain the total $\sum d^2$. Apply, formula (10-4) to calculate correlation coefficient.

Example 10.8: The coefficient of rank correlation between debenture prices and share prices is found to be 0.143. If the sum of the squares of the differences in ranks is given to be 48, find the values of n .

Solution: The formula for Spearman's correlation coefficient is as follows:

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Given, $R = 0.143$, $\sum d^2 = 48$ and $n = 7$. Substituting values in the formula, we get

$$0.143 = 1 - \frac{6 \times 48}{n(n^2 - 1)} = 1 - \frac{288}{n^3 - n}$$

$$0.143(n^3 - n) = (n^3 - n) - 288$$

$$n^3 - n - 336 = 0 \quad \text{or} \quad (n - 7)(n^2 + 7n + 48) = 0$$

This implies that either $n - 7 = 0$, that is, $n = 7$ or $n^2 + 7n + 48 = 0$. But $n^2 + 7n + 48 = 0$ on simplification gives undesirable value of n because its discriminant $b^2 - 4ac$ is negative. Hence $n = 7$.

Example 10.9: The ranks of 15 students in two subjects A and B, are given below. The two numbers within brackets denote the ranks of a student in A and B subjects respectively.

$$(1, 10), (2, 7), (3, 2), (4, 6), (5, 4), (6, 8), (7, 3), (8, 1), \\ (9, 11), (10, 15), (11, 9), (12, 5), (13, 14), (14, 12), (15, 13)$$

Find Spearman's rank correlation coefficient. [Sukhadia Univ., MBA, 1998]

Solution: Since ranks of students with respect to their performance in two subjects are given, calculations for rank correlation coefficient are shown below:

<i>Rank in A</i>	<i>Rank in B</i>	<i>Difference</i>	d^2
R_1	R_2	$d = R_1 - R_2$	
1	10	-9	81
2	7	-5	25
3	2	1	1
4	6	-2	4
5	4	1	1
6	8	-2	4
7	3	4	16
8	1	7	49
9	11	-2	4
10	15	-5	25
11	9	2	4
12	5	7	49
13	14	-1	1
14	12	2	4
15	13	2	4
			$\Sigma d^2 = 272$

$$\text{Apply the formula, } R = 1 - \frac{6 \sum d^2}{n^3 - n} = 1 - \frac{6 \times 272}{15 \{(15)^2 - 1\}} \\ = 1 - \frac{1632}{3360} = 1 - 0.4857 = 0.5143$$

The result shows a moderate degree positive correlation between performance of students in two subjects.

Example 10.10: An office has 12 clerks. The long-serving clerks feel that they should have a seniority increment based on length of service built into their salary structure. An assessment of their efficiency by their departmental manager and the personnel department produces a ranking of efficiency. This is shown below together with a ranking of their length of service.

Ranking according to length of service :	1	2	3	4	5	6	7	8	9	10	11	12
Ranking according to efficiency :	2	3	5	1	9	10	11	12	8	7	6	4

Do the data support the clerks' claim for seniority increment?

[Sukhadia Univ., MBA, 1991]

Solution: Since ranks are already given, calculations for rank correlation coefficient are shown below:

<i>Rank According to Length of Service</i>	<i>Rank According to Efficiency</i>	<i>Difference</i>	d^2
R_1	R_2	$d = R_1 - R_2$	
1	2	-1	1
2	3	-1	1
3	5	-2	4
4	1	3	9
5	9	-4	16
6	10	-4	16
7	11	-4	16
8	12	-4	16
9	8	1	1
10	7	3	9
11	6	5	25
12	4	8	64
			$\Sigma d^2 = 178$

$$\text{Applying the formula, } R = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6 \times 178}{12(144 - 1)} = 1 - \frac{1068}{1716} = 0.378$$

The result shows a low degree positive correlation between length of service and efficiency, the claim of the clerks for a seniority increment based on length of service is not justified.

Example 10.11: Ten competitors in a beauty contest are ranked by three judges in the following order:

Judge 1:	1	6	5	10	3	2	4	9	7	8
Judge 2:	3	5	8	4	7	10	2	1	6	9
Judge 3:	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to determine which pair of judges has the nearest approach to common tastes in beauty. [MD Univ., MBA, 1996]

Solution: The pair of judges who have the nearest approach to common taste in beauty can be obtained in ${}^3C_2 = 3$ ways as follows:

- (i) Judge 1 and judge 2.
- (ii) Judge 2 and judge 3.
- (iii) Judge 3 and judge 1.

Calculations for comparing their ranking are shown below:

Judge 1 R_1	Judge 2 R_2	Judge 3 R_3	$d^2 = (R_1 - R_2)^2$	$d^2 = (R_2 - R_3)^2$	$d^2 = (R_3 - R_1)^2$
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4
3	7	1	16	36	4
2	10	2	64	64	0
4	2	3	4	1	1
9	1	10	64	81	1
7	6	5	1	1	4
8	9	7	1	4	1
$\Sigma d^2 = 200$			$\Sigma d^2 = 214$		$\Sigma d^2 = 60$

Applying the formula

$$R_{12} = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 200}{10(100 - 1)} = 1 - \frac{1200}{990} = -0.212$$

$$R_{23} = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 214}{10(100 - 1)} = 1 - \frac{1284}{990} = -0.297$$

$$R_{13} = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 60}{10(100 - 1)} = 1 - \frac{360}{990} = 0.636$$

Since the correlation coefficient $R_{13} = 0.636$ is largest, the judges 1 and 3 have nearest approach to common tastes in beauty.

Case 2: When Ranks are not Given

When pairs of observations in the data set are not ranked as in Case 1, the ranks are assigned by taking either the highest value or the lowest value as 1 for both the variable's values.

Example 10.12: Quotations of index numbers of security prices of a certain joint stock company are given below:

Year	Debenture Price	Share Price
1	97.8	73.2
2	99.2	85.8
3	98.8	78.9
4	98.3	75.8
5	98.4	77.2
6	96.7	87.2
7	97.1	83.8

Using the rank correlation method, determine the relationship between debenture prices and share prices. [Calicut Univ., BCom, 1997]

Solution: Let us start ranking from the lowest value for both the variables, as shown below:

Debenture Price (x)	Rank	Share Price (y)	Rank	Difference $d = R_1 - R_2$	$d^2 = (R_1 - R_2)^2$
97.8	3	73.2	1	2	4
99.2	7	85.8	6	1	1
98.8	6	78.9	4	2	4
98.3	4	75.8	2	2	4
98.4	5	77.2	3	2	4
96.7	1	87.2	7	-6	36
97.1	2	83.8	5	-3	9
				$\Sigma d^2 = 62$	

$$\begin{aligned} \text{Applying the formula } R &= 1 - \frac{6 \sum d^2}{n^3 - n} = 1 - \frac{6 \times 62}{(7)^3 - 7} \\ &= 1 - \frac{372}{336} = 1 - 0.107 = -0.107 \end{aligned}$$

The result shows a low degree of negative correlation between the debenture prices and share prices of a certain joint stock company.

Example 10.13 An economist wanted to find out if there was any relationship between the unemployment rate in a country and its inflation rate. Data gathered from 7 countries for the year 2004 are given below:

Country	Unemployment Rate (Percent)	Inflation Rate (Per cent)
A	4.0	3.2
B	8.5	8.2
C	5.5	9.4
D	0.8	5.1
E	7.3	10.1
F	5.8	7.8
G	2.1	4.7

Find the degree of linear association between a country's unemployment rate and its level of inflation.

Solution: Let us start ranking from the lowest value for both the variables as shown below:

Unemployment Rate (x)	Rank R_1	Inflation Rate (y)	Rank R_2	Difference $d = R_1 - R_2$	$d^2 = (R_1 - R_2)^2$
4.0	3	3.2	1	2	4
8.5	7	8.2	5	2	4
5.5	4	9.4	6	-2	4
0.8	1	5.1	3	-2	4
7.3	6	10.1	7	-1	1
5.8	5	7.8	4	1	1
2.1	2	4.7	2	0	0
					$\Sigma d^2 = 18$

Applying the formula,

$$R = 1 - \frac{6 \sum d^2}{n^3 - n} = 1 - \frac{6 \times 18}{(7)^3 - (7)} = 1 - \frac{108}{336} = 0.678$$

The result shows a moderately high degree of positive correlation between unemployment rate and inflation rate of seven countries.

Case 3: When Ranks are Equal

While ranking observations in the data set by taking either the highest value or lowest value as rank 1, we may come across a situation of more than one observations being of equal size. In such a case the rank to be assigned to individual observations is an average of the ranks which these individual observations would have got had they differed from each other. For example, if two observations are ranked equal at third place, then the average rank of $(3 + 4)/2 = 3.5$ is assigned to these two observations. Similarly, if three observations are ranked equal at third place, then the average rank of $(3 + 4 + 5)/3 = 4$ is assigned to these three observations.

While equal ranks are assigned to a few observations in the data set, an adjustment is made in the Spearman rank correlation coefficient formula as given below:

$$R = 1 - \frac{6 \left\{ \sum d^2 + \frac{1}{12} (m_1^3 - m_1) + \frac{1}{12} (m_2^3 - m_2) + \dots \right\}}{n(n^2 - 1)}$$

where m_i ($i = 1, 2, 3, \dots$) stands for the number of times an observation is repeated in the data set for both variables.

Example 10.14: A financial analyst wanted to find out whether inventory turnover influences any company's earnings per share (in per cent). A random sample of 7 companies listed in a stock exchange were selected and the following data was recorded for each.

Company	Inventory Turnover	Earnings per Share (Per cent)
	(Number of Times)	
A	4	11
B	5	9
C	7	13
D	8	7
E	6	13
F	3	8
G	5	8

Find the strength of association between inventory turnover and earnings per share. Interpret this finding.

Solution: Let us start ranking from lowest value for both the variables. Since there are tied ranks, the sum of the tied ranks is averaged and assigned to each of the tied observations as shown below.

Inventory Turnover (x)	Rank R_1	Earnings Per Share (y)	Rank R_2	Difference $d = R_1 - R_2$	$d^2 = (R_1 - R_2)^2$
4	2	11	5	-3.0	9.00
5	3.5	9	4	-0.5	0.25
7	6	13	6.5	0.5	0.25
8	7	7	1	6.0	36.00
6	5	13	6.5	-1.5	2.25
3	1	8	2.5	-1.5	2.25
5	3.5	8	2.5	1.0	1.00
					$\Sigma d^2 = 51$

If may be noted that a value 5 of variable x is repeated twice ($m_1 = 2$) and values 8 and 13 of variable y is also reperated twice, so $m_2 = 2$ and $m_3 = 2$. Applying the formula:

$$\begin{aligned}
 R &= 1 - \frac{6 \left\{ \sum d^2 + \frac{1}{12} (m_1^3 - m_1) + \frac{1}{12} (m_2^3 - m_2) + \frac{1}{12} (m_3^3 - m_3) \right\}}{n(n^2 - 1)} \\
 &= 1 - \frac{6 \left\{ 51 + \frac{1}{12} (2^3 - 2) + \frac{1}{12} (2^3 - 2) + \frac{1}{12} (2^3 - 2) \right\}}{7(49 - 1)} \\
 &= 1 - \frac{6 \{ 51 + 0.5 + 0.5 + 0.5 \}}{336} = 1 - 0.9375 = 0.0625
 \end{aligned}$$

The result shows a very week positive association between inventory turnover and earnings per share.

Example 10.15: Obtain the rank correlation coefficient between the variables x and y from the following pairs of observed values.

x : 50 55 65 50 55 60 50 65 70 75

y : 110 110 115 125 140 115 130 120 115 160

[Mangalore Univ., BCom, 1997]

Solution: Let us start ranking from lowest value for both the variables. Moreover, certain observations in both sets of data are repeated, the ranking is done in accordance with suitable average value as shown below.

Variable x	Rank R_1	Variable y	Rank R_2	Difference $d = R_1 - R_2$	$d^2 = (R_1 - R_2)^2$
50	2	110	1.5	0.5	0.25
55	4.5	110	1.5	3.0	9.00
65	7.5	115	4	3.5	12.25
50	2	125	7	-5.0	25.00
55	4.5	140	9	-4.5	20.25
60	6	115	4	2.0	4.00
50	2	130	8	-6.0	36.00
65	7.5	120	6	1.5	2.25
70	9	115	4	5.0	25.00
75	10	160	10	0.0	00.00
					$\Sigma d^2 = 134.00$

It may be noted that for variable x, 50 is repeated thrice ($m_1 = 3$), 55 is repeated twice ($m_2 = 2$), and 65 is repeated twice ($m_3 = 2$). Also for variable y, 110 is repeated twice ($m_4 = 2$) and 115 thrice ($m_5 = 3$). Applying the formula:

$$\begin{aligned}
 R &= 1 - \frac{6 \left\{ \sum d^2 + \frac{1}{12} (m_1^3 - m_1) + \frac{1}{12} (m_2^3 - m_2) + \frac{1}{12} (m_3^3 - m_3) + \frac{1}{12} (m_4^3 - m_4) + \frac{1}{12} (m_5^3 - m_5) \right\}}{n(n^2 - 1)} \\
 &= 1 - \frac{6 \left\{ 134 + \frac{1}{12} (3^3 - 3) + \frac{1}{12} (2^3 - 2) + \frac{1}{12} (2^3 - 2) + \frac{1}{12} (2^3 - 2) + \frac{1}{12} (3^3 - 3) \right\}}{10(100 - 1)} \\
 &= 1 - \frac{6 [134 + 2 + 0.5 + 0.5 + 0.5 + 2]}{990} = 1 - \frac{6 \times 139.5}{990} = 1 - \frac{837}{990} \\
 &= 1 - 0.845 = 0.155
 \end{aligned}$$

The result shows a weak positive association between variables x and y .

10.5.6 Method of Least-Squares

The method of least-squares to calculate the correlation coefficient requires the values of regression coefficients b_{xy} and b_{yx} , so that

$$r = \sqrt{b_{xy} \times b_{yx}}$$

In other words, correlation coefficient is the geometric mean of two regression coefficients (see Chapter 11 for details).

10.5.7 Auto-Correlation Coefficient

The auto correlation coefficient describes the association or mutual dependence between values of the same variable but at different time periods. The auto correlation coefficient provides important information on how a variable relates to itself for a specific time lag. The difference in the period before a cause-and-effect relationship is established is called 'lag'. While computing the correlation, the time gap must be considered, otherwise misleading (deceptive) conclusions may be arrived at. For example, the decrease or increase in supply of a commodity may not immediately reflect on its price, it may take some lead time or time lag.

The formula for auto-correlation coefficient at time lag k is stated as:

$$r_k = \frac{\sum_{i=1}^{n-k} (x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

where k = length of time lag

n = number of observations

\bar{x} = mean of all observations

Example 10.16: The monthly sales of a product, in thousands of units, in the last 6 months are given below:

Month :	1	2	3	4	5	6
Sales :	1.8	2.5	3.1	3.0	4.2	3.4

Compute the auto-correlation coefficient upto lag 2. What conclusion can be derived from these values regarding the presence of a trend in the data?

Solution: The calculations for auto-correlation coefficient are shown below:

Time	Sales (x)	$x_1 = \text{One Time Lag}$	$x_2 = \text{Two Time Lags}$
		Variable Constructed From x	Variable Constructed From x
1	1.8	2.5	3.1
2	2.5	3.1	3.0
3	3.1	3.0	4.2
4	3.0	4.2	3.4
5	4.2	3.4	—
6	3.4	—	—

$$\text{For } k = 1, \bar{x} = \frac{1}{6} (1.8 + 2.5 + \dots + 3.4) = 3$$

$$\{(1.8 - 3)(2.5 - 3) + (2.5 - 3)(3.1 - 3) + (3.1 - 3)(3 - 3) \\ + (3 - 3)(4.2 - 3) + (4.2 - 3)(3.4 - 3)\}$$

$$r_1 = \frac{(1.8 - 3)^2 + (2.5 - 3)^2 + (3.1 - 3)^2 + (3 - 3)^2 + (4.2 - 3)^2 + (3.4 - 3)^2}{(1.8 - 3)^2 + (2.5 - 3)^2 + (3.1 - 3)^2 + (3 - 3)^2 + (4.2 - 3)^2 + (3.4 - 3)^2}$$

$$= \frac{(-1.2)(-0.5) + (-0.5)(0.1) + (0.1)(0) + (0)(1.2) + (1.2)(0.4)}{1.44 + 0.25 + 0.01 + 0 + 1.44 + 0.16}$$

$$= \frac{(0.6 - 0.5 + 0.48)}{3.3} = 0.312$$

For $k = 2$

$$r_2 = \frac{(1.8 - 3)(3.1 - 3) + (2.5 - 3)(3 - 3) + (3.1 - 3)(4.2 - 3) + (3 - 3)(3.4 - 3)}{(1.8 - 3)^2 + (2.5 - 3)^2 + \dots + (3.4 - 3)^2}$$

$$= \frac{(-1.2 \times 0.1) + (-0.5 \times 0) + (0.1 \times 1.2) + (0 \times 0.4)}{3.3} = \frac{-0.12 + 0.12}{3.3} = 0$$

Since the value of r_1 is positive, it implies that there is a seasonal pattern of 6 months duration and $r_2 = 0$ implies that there is no significant change in sales.

Self-Practice Problems 10B

- 10.11** The coefficient of rank correlation of the marks obtained by 10 students in statistics and accountancy was found to be 0.2. It was later discovered that the difference in ranks in two subjects obtained by one of the students was wrongly taken as 9 instead of 7. Find the correct coefficient of rank correlation.

[Delhi Univ., BCom, 1996]

- 10.12** The ranking of 10 students in accordance with their performance in two subjects A and B are as follows:

A:	6	5	3	10	2	4	9	7	8	1
B:	3	8	4	9	1	6	10	7	5	2

Calculate the rank correlation coefficient and comment on its value.

- 10.13** Calculate Spearman's coefficient of correlation between marks assigned to ten students by judges x and y in a certain competitive test as shown below:

Student	Marks by Judge x	Marks by Judge y
1	52	65
2	53	68
3	42	43
4	60	38
5	45	77
6	41	48
7	37	35
8	38	30
9	25	25
10	27	50

- 10.14** An examination of eight applicants for a clerical post was taken by a firm. From the marks obtained by

the applicants in the accountancy and statistics papers, compute the rank correlation coefficient.

Applicant	:	A	B	C	D	E	F	G	H
Marks in accountancy:		15	20	28	12	40	60	20	80
Marks in statistics :		40	30	50	30	20	10	30	60

- 10.15** Seven methods of imparting business education were ranked by the MBA students of two universities as follows:

Method of Teaching	:	1	2	3	4	5	6	7
Rank by students of Univ. A	:	2	1	5	3	4	7	6
Rank by students of Univ. B	:	1	3	2	4	7	5	6

Calculate the rank correlation coefficient and comment on its value.

- 10.16** An investigator collected the following data with respect to the socio-economic status and severity of respiratory illness.

Patient	:	1	2	3	4	5	6	7	8
Socio-economic status (rank)	:	6	7	2	3	5	4	1	8
Severity of illness rank)	:	5	8	4	3	7	1	2	6

Calculate the rank correlation coefficient and comment on its value.

- 10.17** You are given the following data of marks obtained by 11 students in statistics in two tests, one before and other after special coaching:

<i>First Test (Before coaching)</i>	<i>Second Test (After coaching)</i>
23	24
20	19
19	22
21	18
18	20
20	22
18	20
20	22
18	20
17	20
23	23
16	20
19	17

Do the marks indicate that the special coaching has benefited the students? [Delhi Univ., MCom, 1989]

- 10.18** Two departmental managers ranked a few trainees according to their perceived abilities. The ranking are given below:

Trainee :	A	B	C	D	E	F	G	H	I	J
Manager A :	1	9	6	2	5	8	7	3	10	4
Manager B :	3	10	8	1	7	5	6	2	9	4

Calculate an appropriate correlation coefficient to measure the consistency in the ranking.

- 10.19** In an office some keyboard operators, who were already ranked on their speed, were also ranked on accuracy by their supervisor. The results were as follows:

Operator :	A	B	C	D	E	F	G	H	I	J
Speed :	1	2	3	4	5	6	7	8	9	10
Accuracy :	7	9	3	4	1	6	8	2	10	5

Calculate the appropriate correlation coefficient between speed and accuracy.

- 10.20** The personnel department is interested in comparing the ratings of job applicants when measured by a variety of standard tests. The ratings of 9 applicants on interviews and standard psychological test are shown below:

Applicant :	A	B	C	D	E	F	G	H	I
Interview :	5	2	9	4	3	6	1	8	7
Standard test :	8	1	7	5	3	4	2	9	6

Calculate Spearman's rank correlation coefficient and comment on its value.

Hints and Answers

10.11 Given $R = 0.2$, $n = 10$; $R = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$ or

$$0.2 = 1 - \frac{6 \sum d^2}{10(100 - 1)} \text{ or } \sum d^2 = 100$$

Correct value of $R = 1 - \frac{6 \times 100}{10 \times 99} = 0.394$

10.12 $R = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 36}{10(100 - 1)} = 0.782$

10.13 $R = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 76}{10(100 - 1)} = 0.539$

10.14 $R = 1 - \frac{6 \left\{ \sum d^2 + \frac{1}{12} (m_1^3 - m_1) + \frac{1}{12} (m_2^3 - m_2) \right\}}{n(n^2 - 1)}$

$$= 1 - \frac{6 \left\{ 81.5 + \frac{1}{12} (2^3 - 2) + \frac{1}{12} (3^3 - 3) \right\}}{8(64 - 1)} = 0$$

10.15 $R = 0.50$

10.17 $R = 0.71$

10.19 $R = 0.006$

10.16 $R = 0.477$

10.18 $R = 0.842$

10.20 $R = 0.817$

10.6 HYPOTHESIS TESTING FOR CORRELATION COEFFICIENT

We often use the sample correlation coefficient r as an estimator to test whether the possible strength of association between two random variables in the population exist. In other words, we use r as an estimator in testing null and alternative hypotheses about true *population correlation coefficient* ρ (Greek letter rho). When such hypotheses are tested, the assumptions of normal distribution of two random variables, say x and y is required.

10.6.1 Hypothesis Testing about Population Correlation Coefficient (Small Sample)

The test of hypothesis for the existence of a linear relationship between two variables x and y involves the determination of sample correlation coefficient r . This test of linear relationship between x and y is the same as determining whether there is any significant correlation between them. For determining the correlation, we start by hypothesizing the population correlation coefficient ρ equal to zero. The population correlation coefficient ρ measures the degree of association between two variables in a population of interest. The null and alternative hypotheses are expressed as:

- **Two-tailed Test**

$$H_0 : \rho = 0 \text{ (No correlation between variables } x \text{ and } y\text{)}$$

$$H_1 : \rho \neq 0 \text{ (Correlation exists between variables } x \text{ and } y\text{)}$$

- **One-tailed Test**

$$H_0 : \rho = 0 \text{ and } H_1 : \rho > 0 \text{ (or } \rho < 0\text{)}$$

The t -test statistic for testing the null hypothesis is given by:

$$t = \frac{r - \rho}{s_r} = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = \frac{r \times \sqrt{n - 2}}{\sqrt{1 - r^2}}$$

where r = sample correlation coefficient

s_r = standard error of correlation coefficient

n = sample size

The t -test statistic follows t -distribution with $n - 2$ degrees of freedom. If the sample size is large, then the standard error of correlation coefficient is given by $s_r = (1 - r^2)/\sqrt{n}$.

Decision Rule: The calculated value of t -test statistic is compared with its critical (or table) value at $n - 2$ degrees of freedom and level of significance α to arrive at a decision as follows:

One-tailed Test	Two-tailed Test
<ul style="list-style-type: none"> • Reject H_0 if $t_{cal} > t_{\alpha, n-2}$ or $t_{cal} < -t_{\alpha}$ • Otherwise accept H_0 	<ul style="list-style-type: none"> • Reject H_0 if $t_{cal} > t_{\alpha/2, n-2}$ • Otherwise accept H_0

Example 10.17: A random sample of 27 pairs of observations from a normal population gives a correlation coefficient of 0.42. Is it likely that the variables in the population are uncorrelated? [Delhi Univ., MCom. 1997]

Solution: Let us take the null hypothesis that there is no significant difference in the sample and population correlation coefficients, that is,

$$H_0 : \rho = 0 \text{ and } H_1 : \rho \neq 0$$

Given $n = 27$, $df = n - 2 = 25$, $r = 0.42$. Applying t -test statistic to test the null hypothesis, H_0 :

$$\begin{aligned} t &= \frac{r - \rho}{s_r} = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = \frac{0.42}{\sqrt{1 - (0.42)^2}/(27 - 2)} \\ &= \frac{0.42}{0.908/5} = 2.312 \end{aligned}$$

Since the calculated value of $t = 2.312$ is more than the table value $t = 1.708$ at $\alpha = 0.05$ and $df = 25$, the null hypothesis is rejected. Hence it is likely that the variables in the population are not correlated.

Example 10.18: How many pairs of observations must be included in a sample so that an observed correlation coefficient of value 0.42 shall have a calculated value of t greater than 2.72?

Solution: Given, $r = 0.42$, $t = 2.72$. Applying t -test statistic, we get

$$\begin{aligned}\frac{r}{\sqrt{(1-r^2)/(n-2)}} &= t \quad \text{or} \quad r^2 \times \frac{n-2}{1-r^2} = t^2 \\ (0.42)^2 \times \frac{(n-2)}{1-(0.42)^2} &= (2.72)^2 \\ n - 2 &= \frac{(2.72)^2 [1-(0.42)^2]}{(0.42)^2} = \frac{7.3984(0.8236)}{0.1764} \\ &= \frac{6.0933}{0.1764} = 34.542 \\ n &= 2 + 34.542 = 36.542 \cong 37\end{aligned}$$

Hence, the sample size should be of 37 pairs of observations.

Example 10.19: To study the correlation between the stature of father and son, a sample of 1600 is taken from the universe of fathers and sons. The sample study gives the correlation between the two to be 0.80. Within what limits does it hold true for the universe?

Solution: Since the sample size is large, the standard error of the correlation coefficient is given by

$$SE_r = \frac{1-r^2}{\sqrt{n}}$$

Given, correlation coefficient, $r = 0.8$ and $n = 1600$. Thus

$$\text{Standard error } SE_r = \frac{1-(0.8)^2}{\sqrt{1600}} = \frac{1-0.64}{40} = \frac{0.36}{40} = 0.009$$

The limits within which the correlation coefficient should hold true is given by

$$r \pm 3SE_r = 0.80 \pm 3(0.009) \quad \text{or} \quad 0.773 \leq r \leq 0.827$$

10.6.2 Hypothesis Testing about Population Correlation Coefficient (Large Sample)

Since the distribution of sample correlation coefficient r is not normal and its probability curve is skewed in the neighbourhood of population correlation coefficient $\rho = \pm 1$, even for large sample size n , therefore we use Fisher's z -transformation for transforming r into z , using the formula:

$$z = \frac{1}{2} \log_e \frac{1+r}{1-r}$$

The value of z can be seen for different values of r from the standard tables given in the Appendix.

Changing common logarithm to the base e to natural logarithm to the base 10 by multiplying with the constant 2.3026, that is,

$$\log_e x = 2.3026 \log_{10} x$$

where x is a positive integer. Thus the transformation formula becomes

$$z = \frac{1}{2} (2.3026) \log_{10} \frac{1+r}{1-r} = 1.1513 \log_{10} \frac{1+r}{1-r}$$

Fisher showed that the distribution of z is approximately normal with

$$\text{Mean } z_\rho = \frac{1}{2} \log_e \frac{1+\rho}{1-\rho} = 1.1513 \log_{10} \frac{1+\rho}{1-\rho}$$

$$\text{and Standard deviation } \sigma_z = \frac{1}{\sqrt{n-3}}$$

This approximation is useful for large sample sizes, say $n > 50$. However it can also be used for small sample sizes but at least $n \geq 10$.

The Z-test statistic to test the null hypothesis $H_0 : \rho = 0$ and $H_1 : \rho \neq 0$ is given by

$$Z = \frac{z - z_p}{\sigma_z} = \frac{z - z_p}{1/\sqrt{n-3}}$$

where σ is the standard error of Z .

Decision rule

- Accept null hypothesis H_0 if $|Z_{\text{cal}}| < \text{Table value of } Z_{\alpha/2}$
- Otherwise reject H_0

10.6.3 Hypothesis Testing about the Difference between Two Independent Correlation Coefficients

The formula for Z-test statistic given above can be generalized to test the hypothesis of two correlation coefficients r_1 and r_2 derived from two independent samples as follows:

$$Z = \frac{z_1 - z_2}{\sigma_{z_1 - z_2}} = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

where $z_1 = \frac{1}{2} \log_e \frac{1+r_1}{1-r_1} = 1.1513 \log_{10} \frac{1+r_1}{1-r_1}$, and

$$z_2 = \frac{1}{2} \log_e \frac{1+r_2}{1-r_2} = 1.1513 \log_{10} \frac{1+r_2}{1-r_2}$$

are approximately normally distributed with zero mean and unit standard deviation.

The null hypothesis H_0 is accepted if the absolute value $|Z_{\text{cal}}|$ is less than the table value $Z_{\alpha/2}$. Otherwise reject H_0 .

Example 10.20: What is the probability that a correlation coefficient of 0.75 or less arises in a sample of 30 pairs of observations from a normal population in which the true correlation is 0.9?

Solution: Given, $r = 0.75$, $n = 30$, and $\rho = 0.9$. Applying Fisher's z-transformation, we get

$$\begin{aligned} z &= 1.1513 \log_{10} \frac{1+r}{1-r} = 1.1513 \log_{10} \frac{1.75}{0.25} \\ &= 1.1513 [\log_{10} 1.75 - \log_{10} 0.25] \\ &= 1.1513 (0.24304 - 1.39794) = 0.973 \end{aligned}$$

The distribution of z is normal around the true population correlation value $\rho = 0.9$. Therefore

$$\begin{aligned} \text{Mean } z_p &= 1.1513 \log_{10} \frac{1+\rho}{1-\rho} = 1.1513 \log_{10} \frac{1+0.90}{1-0.90} = 1.1513 \log_{10} \frac{1.90}{0.10} \\ &= 1.1513 [\log_{10} 1.90 - \log_{10} 0.10] = 1.1513 (0.27875 + 1) = 1.47 \end{aligned}$$

Thus, the Z-test statistic is given by

$$Z = \frac{|z - z_p|}{\sigma_z} = \frac{|z - z_p|}{1/\sqrt{n-3}} = \frac{|0.973 - 1.47|}{1/\sqrt{30-3}} = 0.498 \times 5.196 = 2.59$$

Hence $P(r \leq 0.75) = P[Z \leq 2.59] = 1 - 0.9952 = 0.0048$.

Example 10.21: Test the significance of the correlation $r = 0.5$ from a sample of size 18 against hypothesized population correlation $\rho = 0.70$.

Solution: Let us take the null hypothesis that the difference is not significant, that is,

$$H_0 : \rho = 0.70 \text{ and } H_1 : \rho \neq 0.70$$

Given $n = 18$, $r = 0.5$. Applying z-transformation, we have

$$\begin{aligned} z &= 1.1513 \log_{10} \frac{1+r}{1-r} = 1.1513 \log_{10} \frac{1+0.5}{1-0.5} \\ &= 1.1513 \log_{10} \frac{1.50}{0.5} = 1.1513 \log_{10} 3 \\ &= 1.1513 (0.4771) = 0.5492 \end{aligned}$$

and Mean $z_p = 1.1513 \log_{10} \frac{1+\rho}{1-\rho} = 1.1513 \log_{10} \frac{1+0.70}{1-0.70}$

$$\begin{aligned} &= 1.1513 \log_{10} \frac{1.70}{0.30} = 1.1513 \log_{10} 5.67 \\ &= 1.1513 (0.7536) = 0.8676 \end{aligned}$$

Applying Z-test statistic, we get

$$\begin{aligned} Z &= \frac{|z - z_p|}{\sigma_z} = \frac{|z - z_p|}{1/\sqrt{n-3}} = |z - z_p| \sqrt{n-3} \\ &= |0.5492 - 0.8676| \sqrt{15} = 0.3184 (3.872) = 1.233 \end{aligned}$$

Since calculated value of $Z_{\text{cal}} = 1.233$ is less than its table value $Z_{\alpha/2} = 1.96$ at 5 per cent significance level, the null hypothesis is accepted. Hence we conclude that the difference (if any) is due to sampling error.

Example 10.22: Two independent samples of size 23 and 21 pairs of observations were analysed and their coefficient of correlation was found as 0.5 and 0.8, respectively. Do these values differ significantly?

Solution: Let us take the null hypothesis that two values do not differ significantly, that is, the samples are drawn from the same population.

Given $n_1 = 23$, $r_1 = 0.5$; $n_2 = 28$, $r_2 = 0.8$. Applying Z-test statistic as follows:

$$\begin{aligned} Z &= \frac{|z_1 - z_2|}{\sigma_{z_1 - z_2}}; & z_1 &= 1.1513 \log_{10} \frac{1+r_1}{1-r_1} = 1.1513 \log_{10} \frac{1+0.5}{1-0.5} \\ &= \frac{|z_1 - z_2|}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} = \frac{|0.55 - 1.10|}{\sqrt{\frac{1}{20} + \frac{1}{25}}} = 1.1513 \log_{10} 3 = 0.55 \\ &= \frac{0.55}{0.30} = 1.833 & z_2 &= 1.1513 \log_{10} \frac{1+r_2}{1-r_2} = 1.1513 \log_{10} \frac{1+0.8}{1-0.8} \\ &&&= 1.1513 \log_{10} 9 = 1.10 \end{aligned}$$

Since the calculated value of $Z_{\text{cal}} = 1.833$ is less than its table value $Z_{\alpha/2} = 1.96$ at 5 per cent significance level, the null hypothesis is accepted. Hence the difference in correlation values is not significant.

Conceptual Questions 10A

1. What is the meaning of the coefficient of correlation?
2. Explain the meaning and significance of the term correlation. [Delhi Univ., MBA, 1995]
3. What is meant by 'correlation'? Distinguish between positive, negative, and zero correlation.
4. What are the numerical limits of r^2 and r ? What does it mean when r equals one? zero? minus one?
5. What is correlation? Clearly explain its role with suitable illustration from simple business problems.

[Delhi Univ., MBA, 1997]

[Ranchi Univ., MBA, 1996]

6. What is the relationship between the coefficient of determination and the coefficient of correlation? How is the coefficient of determination interpreted?
7. Does correlation always signify a cause-and-effect relationship between the variables?
[Osmania Univ., MBA, 1990]
8. What information is provided by the coefficient of correlation of a sample? Why is it necessary to perform a test of a hypothesis for correlation?
9. When the result of a test of correlation is significant, what conclusion is drawn if r is positive? If r is negative?
10. What is the t -statistic that is used in a test for correlation? What is meant by the number of degrees of freedom in a test for correlation and how is it used?
11. What is coefficient of rank correlation? Bring out its usefulness. How does this coefficient differ from the coefficient of correlation? [Delhi Univ., MBA, 2000]
12. What is Spearman's rank correlation coefficient? How does it differ from Karl Pearson's coefficient of correlation?
13. (a) What is a scatter diagram? How do you interpret a scatter diagram?
(b) What is a scatter diagram? How does it help in studying the correlation between two variables, in respect of both its direction and degree?
[Delhi Univ., MBA, 1999]
14. Define correlation coefficient ' r ' and give its limitations. What interpretation would you give if told that the correlation between the number of truck accidents per year and the age of the driver is (-) 0.60 if only drivers with at least one accident are considered?

Self-Practice Problems 10C

- 10.21** The correlation between the price of two commodities x and y in a sample of 60 is 0.68. Could the observed value have arisen

- (a) from an uncorrelated population?
(b) from a population in which true correlation was 0.8?

- 10.22** The following data give sample sizes and correlation coefficients. Test the significance of the difference between two values using Fisher's z -transformation.

Sample Size	Value of r
5	0.870
12	0.560

- 10.23** A company wants to study the relationship between R&D expenditure (in Rs 1000's) and annual profit (in Rs 1000's). The following table presents the information for the last 8 years.

Year	1988	87	86	85	84	83	82	81
R&D expenses :	9	7	5	10	4	5	3	2
Annual profit :	45	42	41	60	30	34	25	20

- (a) Estimate the sample correlation coefficient.

- (b) Test the significance of correlation coefficient at a $\alpha = 5$ per cent level of significance.

- 10.24** Find the least value of r in a sample of 27 pairs from a bivariate normal population at $\alpha = 0.05$ level of significance, where $t_{\alpha} = 0.05 = 2.06$ at $df = 25$.

- 10.25** A small retail business has determined that the correlation coefficient between monthly expenses and profits for the past year, measured at the end of each month, is $r = 0.56$. Assuming that both expenses and profits are approximately normal, test at $\alpha = 0.05$ level of significance the null hypothesis that there is no correlation between them.

- 10.26** The manager of a small shop is hopeful that his sales are rising significantly week by week. Treating the sales for the previous six weeks as a typical example of this rising trend, he recorded them in Rs 1000's and analyzed the results. Has the rise been significant?

Week	1	2	3	4	5	6
Sales	2.69	2.62	2.80	2.70	2.75	2.81

Find the correlation coefficient between sales and week and test it for significance at $\alpha = 0.05$.

Hints and Answers

10.21 (a) $z = 1.1513 \log_{10} \frac{1+r}{1-r} = 1.1513 \log_{10} \frac{1+0.68}{1-0.68}$
 $= 1.1513 \log_{10} \frac{1.68}{0.32} = 0.829$

Standard error, $\sigma_z = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{57}} = 0.13$

Test statistic $Z = \frac{z - z_p}{\sigma_z} = \frac{0.829 - 0}{0.13} = 6.38$

Since deviation of z from z_p is 6 times more than σ_z , the hypothesis is not correct, that is, population is correlated.

$$\begin{aligned}\text{Mean } z_p &= 1.1513 \log_{10} \frac{1+\rho}{1-\rho} \\ &= 1.1513 \log_{10} \frac{1.8}{1.2} = 1.099 \\ \therefore Z &= \frac{|z - z_p|}{\sigma_z} = \frac{|0.829 - 1.099|}{0.13} = 2.08 > 2\end{aligned}$$

times standard error, ρ is likely to be less than 0.8.

10.22 Let H_0 : samples are drawn from the same population.

$$\begin{aligned}z_1 &= 1.1513 \log_{10} \frac{1+r_1}{1-r_1} \\ &= 1.1513 \log \frac{1+0.87}{1-0.87} = 1.333 \\ z_2 &= 1.1513 \log_{10} \frac{1+r_2}{1-r_2} \\ &= 1.1513 \log_{10} \frac{1+0.56}{1-0.56} = 0.633 \\ \sigma_{z_1-z_2} &= \sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}} = \sqrt{\frac{1}{5-3} + \frac{1}{12-3}}\end{aligned}$$

Formulae Used

- Karl Pearson's correlation coefficient

$$r = \frac{\text{Covariance between } x \text{ and } y}{\sigma_x \sigma_y}$$

- Deviation from actual mean

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2} \sqrt{\Sigma(y - \bar{y})^2}}$$

- Deviation from assumed mean

$$r = \frac{n \sum d_x d_y - (\sum d_x)(\sum d_y)}{\sqrt{n \sum d_x^2 - (\sum d_x)^2} \sqrt{n \sum d_y^2 - (\sum d_y)^2}}$$

$$d_x = x - A, d_y = y - B$$

A, B = constants

- Bivariate frequency distribution

$$r = \frac{n \sum f d_x d_y - (\sum f d_x)(\sum f d_y)}{\sqrt{n \sum f d_x^2 - (\sum f d_x)^2} \sqrt{n \sum f d_y^2 - (\sum f d_y)^2}}$$

- Using actual values of x and y

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$$

- Standard error of correlation coefficient, r

$$SE_r = \frac{1-r^2}{\sqrt{n}}$$

$$= 0.782$$

$$Z = \frac{z_1 - z_2}{\sigma_{z_1 - z_2}} = \frac{0.7}{0.782} = 0.895$$

Since the calculated value $Z = 0.895$ is less than its table value $Z_\alpha = 2.58$ at $\alpha = 0.01$ level of significance, H_0 is accepted.

- 10.23** (a) $r = 0.95$ (b) Let $H_0 : r = 0$ and $H_1 : r \neq 0$

$$\begin{aligned}t &= \frac{r}{\sqrt{(1-r^2)/(n-2)}} \\ &= \frac{0.95}{\sqrt{[1-(0.95)^2]/(8-2)}} = 7.512\end{aligned}$$

Since $t_{\text{cal}} = 7.512 > t_{\alpha/2} = 2.447$ for $df = 6$, the H_0 is rejected.

$$\mathbf{10.24} \quad t = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{r \sqrt{27-2}}{\sqrt{1-r}} = \frac{5r}{\sqrt{1-r}} > 2.06$$

or $|r| = 0.381$

- 10.25** $r = 0.560$ and $t_{\text{cal}} = 0.576$, H_0 is rejected.

- 10.26** $r = 0.656$ and $t_{\text{cal}} = 0.729$, H_0 is rejected.

- Probable error of correlation coefficient, r

$$PE_r = 0.6745 \frac{1-r^2}{\sqrt{n}}$$

- Coefficient of determination

$$r^2 = \frac{\text{Explained variance}}{\text{Total variance}} = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

- Spearman's rank correlation coefficient

- Ranks are not equal

$$R = 1 - \frac{6 \sum d^2}{n(n^2-1)}$$

- Ranks are equal

$$R = 1 - \frac{6 \left[\sum d^2 + \frac{1}{12} (m_i^3 - m_i) \right]}{n(n^2-1)}$$

$$t = 1, 2, \dots$$

- Hypothesis testing

- Population correlation coefficient r for a small sample

$$t = \frac{r - \rho}{SE_r} = r \sqrt{\frac{n-2}{1-r^2}}$$

- Population correlation coefficient for a large sample

$$Z = \frac{z - z_p}{\sigma_z} = \frac{z - z_p}{1/\sqrt{n-3}}$$

Review-Self Practice Problems

- 10.27** The following are the monthly figures of the advertising expenditure and sales of a firm. It is generally found that advertising expenditure has its impact on sales generally after 2 months. Allowing for this time lag, calculate the coefficient of correlation.

	Months	Advertising Expenditure	Sales	Months	Advertising Expenditure	Sales
						Expenditure
	Jan.	50	1200	July	140	2400
	Feb.	60	1500	Aug.	160	2600
	March	70	1600	Sep.	170	2800
	April	90	2000	Oct.	190	2900
	May	120	2200	Nov.	200	3100
	June	150	2500	Dec.	250	3900

- 10.28** The coefficient of correlation between two variables x and y is 0.64. Their covariance is 16. The variance of x is 19. Find the standard deviation of y series.

- 10.29** Given $r = 0.8$, $\Sigma xy = 60$, $\sigma_y = 2.5$ and $\Sigma x^2 = 90$, find the number of observations, items. x and y are deviations from arithmetic mean.

[Delhi Univ., BCom, 1998]

- 10.30** Calculate the Karl Pearson's coefficient of correlation between age and playing habits from the data given below. Comment on the value

Age	:	20	21	22	23	24	25
No. of students	:	500	400	300	240	200	160
Regular players	:	400	300	180	96	60	24

[Osmania Univ., MBA, 1998]

- 10.31** A survey regarding income and savings provided the following data:

Income (Rs)	Saving (Rs)			
	500	1000	1500	2000
40,000	8	4	—	—
6000	—	12	24	6
8000	—	9	7	2
10,000	—	—	10	5
12,000	—	—	9	4

Compute Karl Pearson's coefficient of correlation and interpret its value.

[Kurukshetra Univ., MBA, 1997]

- 10.32** A company gives on-the-job training to its salesmen, followed by a test. It is considering whether it should terminate the services of any salesman who does not do well in the test. Following data give the test scores and sales (in 1000 Rs) made by nine salesmen during the last one year

Test scores : 14 19 24 21 26 22 15 20 19

Sales : 31 36 48 37 50 45 33 41 39

Compute the coefficient of correlation between test scores and sales. Does it indicate that termination of the services of salesman with low test scores is justified?

[Madurai Univ., MBA, 1999]

- 10.33** Calculate the coefficient of correlation and its probable error from the following:

S.No.	Subject	Per cent Marks in Final Year Exams	Percent Marks in Sessionals
1	Hindi	75	62
2	English	81	68
3	Physics	70	65
4	Chemistry	76	60
5	Maths	77	69
6	Statistics	81	72
7	Botany	84	76
8	Zoology	75	72

- 10.34** Following figures give the rainfall in inches for the year and the production (in 100's kg) for the Rabi crop and Kharif crops. Calculate Karl Pearson's coefficient of correlation, between rainfall and total production

Rainfall : 20 22 24 26 28 30 32

Rabi production : 15 18 20 32 40 39 40

Kharif production : 15 17 20 18 20 21 15

[Pune Univ., MBA, 1996]

- 10.35** President of a consulting firm is interested in the relationship between environmental work factors and the employees turnover rate. He defines environmental factors as those aspects of a job other than salary and benefits. He visited to similar plants and gave each plant a rating 1 to 25 on its environmental factors. He then obtained each plant's turnover rate (Annual in percentage) examined the relationship.

Environmental

rating : 11 19 7 12 13 10 16 22 14 12

Turnover rate : 6 4 8 3 7 8 3 2 5 6

Compute the correlation coefficient between turnover rate and environmental rating and test it.

[IGNOU, 1996]

- 10.36** Sixteen companies in a state have been ranked according to profit earned during a particular financial year, and the working capital for that year. Calculate the rank correlation coefficient

Company	Rank(Profit)	Rank(Working capital)
A	1	13
B	2	16
C	3	14
D	4	15
E	5	10
F	6	12
G	7	4
H	8	11
I	9	5
J	10	9
K	11	8
L	12	3
M	13	1

N	14	6
O	15	7
P	16	2

10.37 Following are the percentage figures of expenditure incurred on clothing (in Rs 100's) and entertainment (in Rs 100's) by an average working class family in a period of 10 years

Year : 1989 90 91 92 93 94 95 96 97 98

Expenditure
on clothing : 24 27 31 32 20 25 33 30 28 22

Expenditure on
entertainment : 11 8 5 3 13 10 2 7 9 2

Compute Spearman's rank correlation coefficient and comment on the result.

Hints and Answers

10.27 $r = 0.918$

10.28 $r = \frac{\Sigma xy}{n\sigma_x\sigma_y}$; $\sigma_x = \sqrt{9} = 3$;

$$0.64 = 16 \cdot \frac{1}{3\sigma_y} \quad \text{or} \quad \sigma_y = 8.33$$

10.29 $r = \frac{\Sigma xy}{n\sigma_x\sigma_y}$ or $r^2 = \frac{(\Sigma xy)^2}{n^2\sigma_x^2\sigma_y^2}$;

$$(0.8)^2 = \frac{(60)^2}{n^2(90/n) \times 6.25} = \frac{3600}{90n \times 6.25}; \\ n = 10$$

10.30 $r = -0.991$

10.32 $r = 0.947$

10.34 $r = 0.917$

10.36 $R = -0.8176$

10.31 $r = 0.0522$

10.33 $r = 0.623$, $PE_r = 0.146$

10.35 $r = -0.801$

10.37 $R = -0.60$

This page is intentionally left blank.

The cause is hidden, but the result is known.

—Ovid

*I never think of the future,
it comes soon enough.*

—Albert Einstein

Regression Analysis

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- use simple linear regression for building models to business data.
- understand how the method of least squares is used to predict values of a dependent (or response) variable based on the values of an independent (or explanatory) variable.
- measure the variability (residual) of the dependent variable about a straight line (also called regression line) and examine whether regression model fits to the data.

11.1 INTRODUCTION

In Chapter 10 we introduced the concept of statistical relationship between two variables such as: level of sales and amount of advertising; yield of a crop and the amount of fertilizer used; price of a product and its supply, and so on. The relationship between such variables indicate the degree and direction of their association, but fail to answer following question:

- Is there any functional (or algebraic) relationship between two variables? If yes, can it be used to estimate the most likely value of one variable, given the value of other variable?

The statistical technique that expresses the relationship between two or more variables in the form of an equation to estimate the value of a variable, based on the given value of another variable, is called *regression analysis*. The variable whose value is estimated using the algebraic equation is called *dependent (or response) variable* and the variable whose value is used to estimate this value is called *independent (regressor or predictor) variable*. The linear algebraic equation used for expressing a dependent variable in terms of independent variable is called *linear regression equation*.

The term regression was used in 1877 by Sir Francis Galton while studying the relationship between the height of father and sons. He found that though ‘tall father has tall sons’, the average height of sons of tall father is x above the general height, the average height of sons is $2x/3$ above the general height. Such a fall in the average height was described by Galton as ‘regression to mediocrity’. However, the theory of Galton is not universally applicable and the term regression is applied to other types of variables in business and economics. The term regression in the literary sense is also referred as ‘moving backward’.

The basic differences between correlation and regression analysis are summarized as follows:

1. Developing an algebraic equation between two variables from sample data and predicting the value of one variable, given the value of the other variable is referred to as regression analysis, while measuring the strength (or degree) of the relationship between two variables is referred as correlation analysis. The sign of correlation coefficient indicates the nature (direct or inverse) of relationship between two variables, while the absolute value of correlation coefficient indicates the extent of relationship.
2. Correlation analysis determines an association between two variables x and y but not that they have a cause-and-effect relationship. Regression analysis, in contrast to correlation, determines the cause-and-effect relationship between x and y , that is, a change in the value of independent variable x causes a corresponding change (*effect*) in the value of dependent variable y if all other factors that affect y remain unchanged.
3. In linear regression analysis one variable is considered as dependent variable and other as independent variable, while in correlation analysis both variables are considered to be independent.
4. *The coefficient of determination r^2 indicates the proportion of total variance in the dependent variable that is explained or accounted for by the variation in the independent variable.* Since value of r^2 is determined from a sample, its value is subject to sampling error. Even if the value of r^2 is high, the assumption of a linear regression may be incorrect because it may represent a portion of the relationship that actually is in the form of a curve.

11.2 ADVANTAGES OF REGRESSION ANALYSIS

The following are some important advantages of regression analysis:

1. Regression analysis helps in developing a regression equation by which the value of a dependent variable can be estimated given a value of an independent variable.
2. Regression analysis helps to determine standard error of estimate to measure the variability or spread of values of a dependent variable with respect to the regression line. Smaller the variance and error of estimate, the closer the pair of values (x, y) fall about the regression line and better the line fits the data, that is, a good estimate can be made of the value of variable y . When all the points fall on the line, the standard error of estimate equals zero.
3. When the sample size is large ($df \geq 29$), the interval estimation for predicting the value of a dependent variable based on standard error of estimate is considered to be acceptable by changing the values of either x or y . The magnitude of r^2 remains the same regardless of the values of the two variables.

11.3 TYPES OF REGRESSION MODELS

The primary objective of regression analysis is the development of a *regression model* to explain the association between two or more variables in the given population. A regression model is the mathematical equation that provides prediction of value of dependent variable based on the known values of one or more independent variables.

The particular form of regression model depends upon the nature of the problem under study and the type of data available. However, each type of association or relationship can be described by an equation relating a dependent variable to one or more independent variables.

11.3.1 Simple and Multiple Regression Models

If a regression model characterizes the relationship between a dependent y and only one independent variable x , then such a regression model is called a *simple regression model*. But if more than one independent variables are associated with a dependent variable,

then such a regression model is called a *multiple regression model*. For example, sales turnover of a product (a dependent variable) is associated with multiple independent variables such as price of the product, expenditure on advertisement, quality of the product, competitors, and so on. Now if we want to estimate possible sales turnover with respect to only one of these independent variables, then it is an example of a simple regression model, otherwise multiple regression model is applicable.

11.3.2 Linear and Nonlinear Regression Models

If the value of a dependent (response) variable y in a regression model tends to increase in direct proportion to an increase in the values of independent (predictor) variable x , then such a regression model is called a *linear model*. Thus, it can be assumed that the mean value of the variable y for a given value of x is related by a straight-line relationship. Such a relationship is called *simple linear regression model* expressed with respect to the population parameters β_0 and β_1 as:

$$E(y|x) = \beta_0 + \beta_1 x \quad (11-1)$$

where β_0 = y -intercept that represents mean (or average) value of the dependent variable y when $x = 0$

β_1 = slope of the regression line that represents the expected change in the value of y (either positive or negative) for a unit change in the value of x .

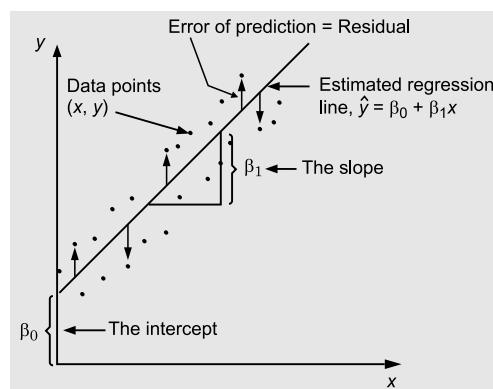


Figure 11.1
Straight Line Relationship

The intercept β_0 and the slope β_1 are *unknown regression coefficients*. The equation (11-1) requires to compute the values of β_0 and β_1 to predict average values of y for a given value of x . However Fig. 11.1 presents a scatter diagram where each pair of values (x_i, y_i) represents a point in a two-dimensional coordinate system. Although the mean or average value of y is a linear function of x , but not all values of y fall exactly on the straight line rather fall around the line.

Since few points do not fall on the regression line, therefore values of y are not exactly equal to the values yielded by the equation: $E(y|x) = \beta_0 + \beta_1 x$, also called *line of mean deviations of observed y value from the regression line*. This situation is responsible for *random error* (also called *residual variation* or *residual error*) in the prediction of y values for given values of x . In such a situation, it is likely that the variable x does not explain all the variability of the variable y . For instance, sales volume is related to advertising, but if other factors related to sales are ignored, then a regression equation to predict the sales volume (y) by using annual budget of advertising (x) as a predictor will probably involve some error. Thus for a fixed value of x , the actual value of y is determined by the *mean value function plus a random error term as follows*:

$$\begin{aligned} y &= \text{Mean value function} + \text{Deviation} \\ &= \beta_0 + \beta_1 x + e = E(y) + e \end{aligned} \quad (11-2)$$

where e is the *observed random error*. This equation is also called *simple probabilistic linear regression model*.

The error component e allows each individual value of y to deviate from the line of means by a small amount. The random errors corresponding to different observations (x_i, y_i) for $i=1, 2, \dots, n$ are assumed to follow a normal distribution with mean zero and (unknown) constant standard deviation.

The term e in the expression (11-2) is called the *random error* because its value, associated with each value of variable y , is assumed to vary unpredictably. The extent of this error for a given value of x is measured by the error variance σ_e^2 . Lower the value of σ_e^2 , better is the fit of linear regression model to a sample data.

If the line passing through the pair of values of variables x and y is curvilinear, then the relationship is called *nonlinear*. A nonlinear relationship implies a varying absolute change in the dependent variable with respect to changes in the value of the independent variable. A nonlinear relationship is not very useful for predictions.

In this chapter, we shall discuss methods of simple linear regression analysis involving single independent variable.

11.4 ESTIMATION : THE METHOD OF LEAST SQUARES

To estimate the values of regression coefficients β_0 and β_1 , suppose a sample of n pairs of observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ is drawn from the population under study. A method that provides the best linear unbiased estimates of β_0 and β_1 is called the *method of least squares*. The estimates of β_0 and β_1 should result in a straight line that is 'best fit' to the data points. The straight line so drawn is referred to as '*best fitted*' (*least squares or estimated*) regression line because the sum of the squares of the vertical deviations (difference between the actual values of y and the estimated values \hat{y} predicted from the fitted line) is as small as possible.

Using equation (11-2), we may express given n observations in the sample data as:

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad \text{or} \quad e_i = y_i - (\beta_0 + \beta_1 x_i), \text{ for all } i$$

Mathematically, we intend to minimize

$$L = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n \{y_i - (\beta_0 + \beta_1 x_i)\}^2$$

Let b_0 and b_1 be the least-squares estimators of β_0 and β_1 respectively. The least-squares estimators b_0 and b_1 must satisfy

$$\begin{aligned} \frac{\partial L}{\partial \beta_0} \Bigg|_{b_0, b_1} &= -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) = 0 \\ \frac{\partial L}{\partial \beta_1} \Bigg|_{b_0, b_1} &= -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) x_i = 0 \end{aligned}$$

After simplifying these two equations, we get

$$\sum_{i=1}^n y_i = nb_0 + b_1 \sum_{i=1}^n x_i \tag{11-3}$$

$$\sum_{i=1}^n x_i y_i = b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2$$

Equations (11-3) are called the *least-squares normal equations*. The values of least squares estimators b_0 and b_1 can be obtained by solving equations (11-3). Hence the *fitted or estimated regression line* is given by:

$$\hat{y} = b_0 + b_1 x$$

where \hat{y} (called y hat) is the value of y lying on the fitted regression line for a given x value and $e_i = y_i - \hat{y}_i$ is called the *residual* that describes the error in fitting of the regression line to the observation y_i . The fitted value \hat{y} is also called the *predicted value of y* because if actual value of y is not known, then it would be predicted for a given value of x using the estimated regression line.

Remark: The sum of the residuals is zero for any least-squares regression line. Since $\sum y_i = \sum \hat{y}_i$, therefore so $\sum e_i = 0$.

11.5 ASSUMPTIONS FOR A SIMPLE LINEAR REGRESSION MODEL

To make valid statistical inference using regression analysis, we make certain assumptions about the bivariate population from which a sample of paired observations is drawn and the manner in which observations are generated. These assumptions form the basis for application of simple linear regression models. Figure 11.2 illustrates these assumptions.

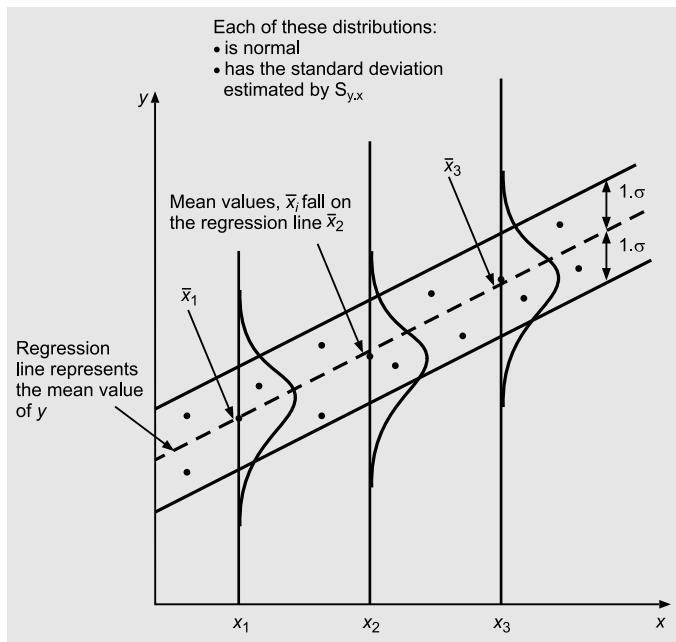


Figure 11.2
 Graphical Illustration of Assumptions in Regression Analysis

Assumptions

1. The relationship between the dependent variable y and independent variable x exists and is linear. The average relationship between x and y can be described by a simple linear regression equation $y = a + bx + e$, where e is the deviation of a particular value of y from its expected value for a given value of independent variable x .
2. For every value of the independent variable x , there is an expected (or mean) value of the dependent variable y and these values are normally distributed. The mean of these normally distributed values fall on the line of regression.
3. The dependent variable y is a continuous random variable, whereas values of the independent variable x are fixed values and are not random.
4. The sampling error associated with the expected value of the dependent variable y is assumed to be an independent random variable distributed normally with mean zero and constant standard deviation. The errors are not related with each other in successive observations.
5. The standard deviation and variance of expected values of the dependent variable y about the regression line are constant for all values of the independent variable x within the range of the sample data.
6. The value of the dependent variable cannot be estimated for a value of an independent variable lying outside the range of values in the sample data.

11.6 PARAMETERS OF SIMPLE LINEAR REGRESSION MODEL

The fundamental aim of regression analysis is to determine a regression equation (line) that makes sense and fits the representative data such that the error of variance is as small as possible. This implies that the regression equation should adequately be used for prediction. J. R. Stockton stated that

- The device used for estimating the values of one variable from the value of the other consists of a line through the points, drawn in such a manner as to represent the average relationship between the two variables. Such a line is called line of regression.

The two variables x and y which are correlated can be expressed in terms of each other in the form of straight line equations called *regression equations*. Such lines should be able to provide the best fit of sample data to the population data. The algebraic expression of regression lines is written as:

- The regression equation of y on x

$$y = a + bx$$

is used for estimating the value of y for given values of x .

- Regression equation of x on y

$$x = c + dy$$

is used for estimating the value of x for given values of y .

Remarks

1. When variables x and y are correlated perfectly (either positive or negative) these lines coincide, that is, we have only one line.
2. Higher the degree of correlation, nearer the two regression lines are to each other.
3. Lesser the degree of correlation, more the two regression lines are away from each other. That is, when $r = 0$, the two lines are at right angle to each other.
4. Two linear regression lines intersect each other at the point of the average value of variables x and y .

11.6.1 Regression Coefficients

To estimate values of population parameter β_0 and β_1 , under certain assumptions, the fitted or estimated regression equation representing the straight line regression model is written as:

$$\hat{y} = a + bx$$

where \hat{y} = estimated average (mean) value of dependent variable y for a given value of independent variable x .

a or b_0 = y -intercept that represents average value of \hat{y}

b = slope of regression line that represents the expected change in the value of y for unit change in the value of x

To determine the value of \hat{y} for a given value of x , this equation requires the determination of two unknown constants a (intercept) and b (also called regression coefficient). Once these constants are calculated, the regression line can be used to compute an estimated value of the dependent variable y for a given value of independent variable x .

The particular values of a and b define a specific linear relationship between x and y based on sample data. The coefficient ' a ' represents the *level of fitted line* (i.e., the distance of the line above or below the origin) when x equals zero, whereas coefficient ' b ' represents the *slope of the line* (a measure of the change in the estimated value of y for a one-unit change in x).

The regression coefficient ' b ' is also denoted as:

- b_{yx} (*regression coefficient of y on x*) in the regression line, $y = a + bx$
- b_{xy} (*regression coefficient of x on y*) in the regression line, $x = c + dy$

Properties of regression coefficients

1. The correlation coefficient is the geometric mean of two regression coefficients, that is, $r = \sqrt{b_{yx} \times b_{xy}}$.
2. If one regression coefficient is greater than one, then other regression coefficient must be less than one, because the value of correlation coefficient r cannot exceed one. However, both the regression coefficients may be less than one.
3. Both regression coefficients must have the same sign (either positive or negative). This property rules out the case of opposite sign of two regression coefficients.

4. The correlation coefficient will have the same sign (either positive or negative) as that of the two regression coefficients. For example, if $b_{yx} = -0.664$ and $b_{xy} = -0.234$, then $r = -\sqrt{0.664 \times 0.234} = -0.394$.
5. The arithmetic mean of regression coefficients b_{xy} and b_{yx} is more than or equal to the correlation coefficient r , that is, $(b_{yx} + b_{xy})/2 \geq r$. For example, if $b_{yx} = -0.664$ and $b_{xy} = -0.234$, then the arithmetic mean of these two values is $(-0.664 - 0.234)/2 = -0.449$, and this value is more than the value of $r = -0.394$.
6. Regression coefficients are independent of origin but not of scale.

11.7 METHODS TO DETERMINE REGRESSION COEFFICIENTS

Following are the methods to determine the parameters of a fitted regression equation.

11.7.1 Least Squares Normal Equations

Let $\hat{y} = a + bx$ be the least squares line of y on x , where \hat{y} is the estimated average value of dependent variable y . The line that minimizes the sum of squares of the deviations of the observed values of y from those predicted is the best fitting line. Thus the sum of residuals for any least-square line is minimum, where

$$L = \sum (y - \hat{y})^2 = \sum \{y - (a + bx)\}^2; \quad a, b = \text{constants}$$

Differentiating L with respect to a and b and equating to zero, we have

$$\frac{\partial L}{\partial a} = -2 \sum \{y - (a + bx)\} = 0$$

$$\frac{\partial L}{\partial b} = -2 \sum \{y - (a + bx)\}x = 0$$

Solving these two equations, we get the same set of equations as equations (11-3)

$$\begin{aligned} \Sigma y &= na + b\Sigma x \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 \end{aligned} \tag{11-4}$$

where n is the total number of pairs of values of x and y in a sample data. The equations (11-4) are called *normal equations* with respect to the regression line of y on x . After solving these equations for a and b , the values of a and b are substituted in the regression equation, $y = a + bx$.

Similarly if we have a least squares line $\hat{x} = c + dy$ of x on y , where \hat{x} is the estimated mean value of dependent variable x , then the normal equations will be

$$\begin{aligned} \Sigma x &= nc + d\Sigma y \\ \Sigma xy &= n\Sigma y + d\Sigma y^2 \end{aligned}$$

These equations are solved in the same manner as described above for constants c and d . The values of these constants are substituted to the regression equation $x = c + dy$.

Alternative method to calculate value of constants

Instead of using the algebraic method to calculate values of a and b , we may directly use the results of the solutions of these normal equation.

The gradient ' b ' (regression coefficient of y on x) and ' d ' (regression coefficient of x on y) are calculated as:

$$b = \frac{S_{xy}}{S_{xx}}, \quad \text{where} \quad S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2$$

$$\text{and } d = \frac{S_{yx}}{S_{yy}}, \quad \text{where} \quad S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2$$

Since the regression line passes through the point (\bar{x}, \bar{y}) , the mean values of x and y and the regression equations can be used to find the value of constants a and c as follows:

$$a = \bar{y} - b\bar{x} \text{ for regression equation of } y \text{ on } x$$

$$c = \bar{x} - d\bar{y} \text{ for regression equation of } x \text{ on } y$$

The calculated values of a , b and c , d are substituted in the regression line $y = a + bx$ and $x = c + dy$ respectively to determine the exact relationship.

Example 11.1: Use least squares regression line to estimate the increase in sales revenue expected from an increase of 7.5 per cent in advertising expenditure.

Firm	Annual Percentage Increase in Advertising Expenditure	Annual Percentage Increase in Sales Revenue
A	1	1
B	3	2
C	4	2
D	6	4
E	8	6
F	9	8
G	11	8
H	14	9

Solution: Assume sales revenue (y) is dependent on advertising expenditure (x). Calculations for regression line using following normal equations are shown in Table 11.1

$$\Sigma y = na + b\Sigma x \quad \text{and} \quad \Sigma xy = a\Sigma x + b\Sigma x^2$$

Table 11.1: Calculation for Normal Equations

Sales Revenue y	Advertising Expenditure, x	x^2	xy
1	1	1	1
2	3	9	6
2	4	16	8
4	6	36	24
6	8	64	48
8	9	81	72
8	11	121	88
9	14	196	126
40	56	524	373

Approach 1 (Normal Equations):

$$\Sigma y = na + b\Sigma x \quad \text{or} \quad 40 = 8a + 56b$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \text{or} \quad 373 = 56a + 524b$$

Solving these equations, we get

$$a = 0.072 \text{ and } b = 0.704$$

Substituting these values in the regression equation

$$y = a + bx = 0.072 + 0.704x$$

For $x = 7.5\%$ or 0.075 increase in advertising expenditure, the estimated increase in sales revenue will be

$$y = 0.072 + 0.704 (0.075) = 0.1248 \text{ or } 12.48\%$$

Approach 2 (Short-cut method):

$$b = \frac{S_{xy}}{S_{xx}} = \frac{93}{132} = 0.704,$$

$$\text{where } S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 373 - \frac{40 \times 56}{8} = 93$$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 524 - \frac{(56)^2}{8} = 132$$

The intercept 'a' on the y-axis is calculated as:

$$a = \bar{y} - b\bar{x} = \frac{40}{8} - 0.704 \times \frac{56}{8} = 5 - 0.704 \times 7 = 0.072$$

Substituting the values of $a = 0.072$ and $b = 0.704$ in the regression equation, we get

$$y = a + bx = 0.072 + 0.704x$$

For $x = 0.075$, we have $y = 0.072 + 0.704(0.075) = 0.1248$ or 12.48%.

Example 11.2: The owner of a small garment shop is hopeful that his sales are rising significantly week by week. Treating the sales for the previous six weeks as a typical example of this rising trend, he recorded them in Rs 1000's and analysed the results

Week :	1	2	3	4	5	6
Sales :	2.69	2.62	2.80	2.70	2.75	2.81

Fit a linear regression equation to suggest to him the weekly rate at which his sales are rising and use this equation to estimate expected sales for the 7th week.

Solution: Assume sales (y) is dependent on weeks (x). Then the normal equations for regression equation: $y = a + bx$ are written as:

$$\Sigma y = na + b\Sigma x \quad \text{and} \quad \Sigma xy = a\Sigma x + b\Sigma x^2$$

Calculations for sales during various weeks are shown in Table 11.2.

Table 11.2: Calculations of Normal Equations

Week (x)	Sales (y)	x^2	xy
1	2.69	1	2.69
2	2.62	4	5.24
3	2.80	9	8.40
4	2.70	16	10.80
5	2.75	25	13.75
6	2.81	36	16.86
21	16.37	91	57.74

The gradient 'b' is calculated as:

$$b = \frac{S_{xy}}{S_{xx}} = \frac{0.445}{17.5} = 0.025; \quad S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 57.74 - \frac{21 \times 16.37}{6} = 0.445$$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 91 - \frac{(21)^2}{6} = 17.5$$

The intercept 'a' on the y-axis is calculated as

$$\begin{aligned} a &= \bar{y} - b\bar{x} = \frac{16.37}{6} - 0.025 \times \frac{21}{6} \\ &= 2.728 - 0.025 \times 3.5 = 2.64 \end{aligned}$$

Substituting the values $a = 2.64$ and $b = 0.025$ in the regression equation, we have

$$y = a + bx = 2.64 + 0.025x$$

For $x = 7$, we have $y = 2.64 + 0.025(7) = 2.815$

Hence the expected sales during the 7th week is likely to be Rs 2.815 (in Rs 1000's).

11.7.2 Deviations Method

Calculations to least squares normal equations become lengthy and tedious when values of x and y are large. Thus the following two methods may be used to reduce the computational time.

(a) **Deviations Taken from Actual Mean Values of x and y** If deviations of actual values of variables x and y are taken from their mean values \bar{x} and \bar{y} , then the regression equations can be written as:

- Regression equation of y on x

$$y - \bar{y} = b_{yx} (x - \bar{x})$$

where b_{yx} = regression coefficient of y on x .

The value of b_{yx} can be calculated using the formula

$$b_{yx} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

- Regression equation of x on y

$$x - \bar{x} = b_{xy} (y - \bar{y})$$

where b_{xy} = regression coefficient of x on y .

The value of b_{xy} can be calculated formula

$$b_{xy} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2}$$

(b) **Deviations Taken from Assumed Mean Values for x and y** If mean value of either x or y or both are in fractions, then we must prefer to take deviations of actual values of variables x and y from their assumed means.

- Regression equation of y on x

$$y - \bar{y} = b_{yx} (x - \bar{x})$$

$$\text{where } b_{yx} = \frac{n \sum d_x d_y - (\sum d_x)(\sum d_y)}{n \sum d_x^2 - (\sum d_x)^2}$$

n = number of observations

$d_x = x - A$; A is assumed mean of x

$d_y = y - B$; B is assumed mean of y

- Regression equation of x on y

$$x - \bar{x} = b_{xy} (y - \bar{y})$$

$$\text{where } b_{xy} = \frac{n \sum d_x d_y - (\sum d_x)(\sum d_y)}{n \sum d_y^2 - (\sum d_y)^2}$$

n = number of observations

$d_x = x - A$; A is assumed mean of x

$d_y = y - B$; B is assumed mean of y

(c) **Regression Coefficients in Terms of Correlation Coefficient** If deviations are taken from actual mean values, then the values of regression coefficients can be alternatively calculated as follows:

$$b_{yx} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

$$= \frac{\text{Covariance}(x, y)}{\sigma_x^2} = r \cdot \frac{\sigma_y}{\sigma_x}$$

$$b_{xy} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2}$$

$$= \frac{\text{Covariance}(x, y)}{\sigma_y^2} = r \cdot \frac{\sigma_x}{\sigma_y}$$

Example 11.3: The following data relate to the scores obtained by 9 salesmen of a company in an intelligence test and their weekly sales (in Rs 1000's)

Salesmen	: A	B	C	D	E	F	G	H	I
Test scores	: 50	60	50	60	80	50	80	40	70
Weekly sales	: 30	60	40	50	60	30	70	50	60

- Obtain the regression equation of sales on intelligence test scores of the salesmen.
- If the intelligence test score of a salesman is 65, what would be his expected weekly sales.
[HP Univ., MCom, 1996]

Solution: Assume weekly sales (y) as dependent variable and test scores (x) as independent variable. Calculations for the following regression equation are shown in Table 11.3.

$$y - \bar{y} = b_{yx} (x - \bar{x})$$

Table 11.3: Calculation for Regression Equation

<i>Weekly Sales, x</i>	$dx = x - 60$	d_x^2	<i>Test Score, y</i>	$dy = y - 50$	d_y^2	$d_x d_y$
50	-10	100	30	-20	400	200
60	0	0	60	10	100	0
50	-10	100	40	-10	100	100
60	0	0	50	0	0	0
80	20	400	60	10	100	200
50	-10	100	30	-20	400	200
80	20	400	70	20	400	400
40	-20	400	50	0	0	0
70	10	100	60	10	100	100
540	0	1600	450	0	1600	1200

$$(a) \bar{x} = \frac{\Sigma x}{n} = \frac{540}{9} = 60; \bar{y} = \frac{\Sigma y}{n} = \frac{450}{9} = 50$$

$$b_{yx} = \frac{\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{1200}{1600} = 0.75$$

Substituting values in the regression equation, we have

$$y - 50 = 0.75(x - 60) \text{ or } y = 5 + 0.75x$$

For test score $x = 65$ of salesman, we have

$$y = 5 + 0.75(65) = 53.75$$

Hence we conclude that the weekly sales is expected to be Rs 53.75 (in Rs 1000's) for a test score of 65.

Example 11.4: A company is introducing a job evaluation scheme in which all jobs are graded by points for skill, responsibility, and so on. Monthly pay scales (Rs in 1000's) are then drawn up according to the number of points allocated and other factors such as experience and local conditions. To date the company has applied this scheme to 9 jobs:

Job :	A	B	C	D	E	F	G	H	I
Points :	5	25	7	19	10	12	15	28	16
Pay (Rs) :	3.0	5.0	3.25	6.5	5.5	5.6	6.0	7.2	6.1

(a) Find the least squares regression line for linking pay scales to points.

(b) Estimate the monthly pay for a job graded by 20 points.

Solution: Assume monthly pay (y) as the dependent variable and job grade points (x) as the independent variable. Calculations for the following regression equation are shown in Table 11.4.

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

Table 11.4: Calculations for Regression Equation

<i>Grade Points, x</i>	$d_x = x - 15$	d_x^2	<i>Pay Scale, y</i>	$d_y = y - 5$	d_y^2	$d_x d_y$
5	-10	100	3.0	-2.0	4	20
25	10	100	(5.0) \leftarrow B	0	0	0
7	-8	64	3.25	-1.75	3.06	14
19	4	16	6.5	1.50	2.25	6
10	-5	25	5.5	0.50	0.25	-2.5
12	-3	9	5.6	0.60	0.36	-1.8
(15) \leftarrow A	0	0	6.0	1.00	1.00	0
28	13	169	7.2	2.2	4.84	28.6
16	1	1	6.1	1.1	1.21	1.1
137	2	484	48.15	3.15	16.97	65.40

$$(a) \bar{x} = \frac{\Sigma x}{n} = \frac{137}{9} = 15.22; \bar{y} = \frac{\Sigma y}{n} = \frac{48.15}{9} = 5.35$$

Since mean values \bar{x} and \bar{y} are non-integer value, therefore deviations are taken from assumed mean as shown in Table 11.4.

$$b_{yx} = \frac{n \sum d_x d_y - (\sum d_x)(\sum d_y)}{n \sum d_x^2 - (\sum d_x)^2} = \frac{9 \times 65.40 - 2 \times 3.15}{9 \times 484 - (2)^2} = \frac{582.3}{4352} = 0.133$$

Substituting values in the regression equation, we have

$$y - \bar{y} = b_{yx} (x - \bar{x}) \text{ or } y - 5.35 = 0.133 (x - 15.22) = 3.326 + 0.133x$$

(b) For job grade point $x=20$, the estimated average pay scale is given by

$$y = 3.326 + 0.133x = 3.326 + 0.133 (20) = 5.986$$

Hence, likely monthly pay for a job with grade points 20 is Rs 5986.

Example 11.5: The following data give the ages and blood pressure of 10 women.

Age : 56 42 36 47 49 42 60 72 63 55

Blood pressure : 147 125 118 128 145 140 155 160 149 150

(a) Find the correlation coefficient between age and blood pressure.

(b) Determine the least squares regression equation of blood pressure on age.

(c) Estimate the blood pressure of a woman whose age is 45 years.

[Ranchi Univ. MBA; South Gujarat Univ., MBA, 1997]

Solution: Assume blood pressure (y) as the dependent variable and age (x) as the independent variable. Calculations for regression equation of blood pressure on age are shown in Table 11.5.

Table 11.5: Calculations for Regression Equation

Age, x	$d_x = x - 49$	d_x^2	Blood, y	$d_y = y - 145$	d_y^2	$d_x d_y$
56	7	49	147	2	4	14
42	-7	49	125	-20	400	140
36	-13	169	118	-27	729	351
47	-2	4	128	-17	289	34
49 ← A	0	0	145 ← B	0	0	0
42	-7	49	140	-5	25	35
60	11	121	155	10	100	110
72	23	529	160	15	225	345
63	14	196	149	4	16	56
55	6	36	150	5	25	30
522	32	1202	1417	-33	1813	1115

(a) Coefficient of correlation between age and blood pressure is given by

$$\begin{aligned} r &= \frac{n \sum d_x d_y - \sum d_x \sum d_y}{\sqrt{n \sum d_x^2 - (\sum d_x)^2} \sqrt{n \sum d_y^2 - (\sum d_y)^2}} \\ &= \frac{10(1115) - (32)(-33)}{\sqrt{10(1202) - (32)^2} \sqrt{10(1813) - (-33)^2}} \\ &= \frac{11150 + 1056}{\sqrt{12020 - 1024} \sqrt{18130 - 1089}} = \frac{12206}{13689} = 0.892 \end{aligned}$$

We may conclude that there is a high degree of positive correlation between age and blood pressure.

(b) The regression equation of blood pressure on age is given by

$$y - \bar{y} = b_{yx} (x - \bar{x})$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{522}{10} = 52.2; \quad \bar{y} = \frac{\Sigma y}{n} = \frac{1417}{10} = 141.7$$

and $b_{yx} = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{10(1115) - 32(-33)}{10(1202) - (32)^2} = \frac{12206}{10996} = 1.11$

Substituting these values in the above equation, we have

$$y - 141.7 = 1.11(x - 52.2) \text{ or } y = 83.758 + 1.11x$$

This is the required regression equation of y on x .

- (c) For a women whose age is 45, the estimated average blood pressure will be

$$y = 83.758 + 1.11(45) = 83.758 + 49.95 = 133.708$$

Hence, the likely blood pressure of a woman of 45 years is 134.

Example 11.6: The General Sales Manager of Kiran Enterprises—an enterprise dealing in the sale of readymade men's wear—is toying with the idea of increasing his sales to Rs 80,000. On checking the records of sales during the last 10 years, it was found that the annual sale proceeds and advertisement expenditure were highly correlated to the extent of 0.8. It was further noted that the annual average sale has been Rs 45,000 and annual average advertisement expenditure Rs 30,000, with a variance of Rs 1600 and Rs 625 in advertisement expenditure respectively.

In view of the above, how much expenditure on advertisement would you suggest the General Sales Manager of the enterprise to incur to meet his target of sales?

[Kurukshetra Univ., MBA, 1998]

Solution: Assume advertisement expenditure (y) as the dependent variable and sales (x) as the independent variable. Then the regression equation advertisement expenditure on sales is given by

$$(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

Given $r = 0.8$, $\sigma_x = 40$, $\sigma_y = 25$, $\bar{x} = 45,000$, $\bar{y} = 30,000$. Substituting these value in the above equation, we have

$$(y - 30,000) = 0.8 \frac{25}{40} (x - 45,000) = 0.5 (x - 45,000)$$

$$y = 30,000 + 0.5x - 22,500 = 7500 + 0.5x$$

When a sales target is fixed at $x = 80,000$, the estimated amount likely to be spent on advertisement would be

$$y = 7500 + 0.5 \times 80,000 = 7500 + 40,000 = \text{Rs } 47,500$$

Example 11.7: You are given the following information about advertising expenditure and sales:

	Advertisement (x) (Rs in lakh)	Sales (y) (Rs in lakh)
Arithmetic mean, \bar{x}	10	90
Standard deviation, σ	3	12

Correlation coefficient = 0.8

- (a) Obtain the two regression equations.
- (b) Find the likely sales when advertisement budget is Rs 15 lakh.
- (c) What should be the advertisement budget if the company wants to attain sales target of Rs 120 lakh? [Kumaon Univ., MBA, 2000, MBA, Delhi Univ., 2002]

Solution: (a) Regression equation of x on y is given by

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

Given $\bar{x} = 10$, $r = 0.8$, $\sigma_x = 3$, $\sigma_y = 12$, $\bar{y} = 90$. Substituting these values in the above regression equation, we have

$$x - 10 = 0.8 \frac{3}{12} (y - 90) \quad \text{or} \quad x = -8 + 0.2y$$

Regression equation of y on x is given by

$$(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$y - 90 = 0.8 \frac{12}{3} (x - 10) \quad \text{or} \quad y = 58 + 3.2x$$

(b) Substituting $x = 15$ in regression equation of y on x . The likely average sales volume would be

$$y = 58 + 3.2 (15) = 58 + 48 = 106$$

Thus the likely sales for advertisement budget of Rs 15 lakh is Rs 106 lakh.

(c) Substituting $y = 120$ in the regression equation of x on y . The likely advertisement budget to attain desired sales target of Rs 120 lakh would be

$$x = -8 + 0.2y = -8 + 0.2 (120) = 16$$

Hence, the likely advertisement budget of Rs 16 lakh should be sufficient to attain the sales target of Rs 120 lakh.

Example 11.8: In a partially destroyed laboratory record of an analysis of regression data, the following results only are legible:

Variance of $x = 9$

Regression equations : $8x - 10y + 66 = 0$ and $40x - 18y = 214$

Find on the basis of the above information:

(a) The mean values of x and y ,

(b) Coefficient of correlation between x and y , and

(c) Standard deviation of y . [Pune Univ., MBA, 1996; CA May 1999]

Solution: (a) Since two regression lines always intersect at a point (\bar{x}, \bar{y}) representing mean values of the variables involved, solving given regression equations to get the mean values \bar{x} and \bar{y} as shown below:

$$8x - 10y = -66$$

$$40x - 18y = 214$$

Multiplying the first equation by 5 and subtracting from the second, we have

$$32y = 544 \quad \text{or} \quad y = 17, \text{i.e. } \bar{y} = 17$$

Substituting the value of y in the first equation, we get

$$8x - 10(17) = -66 \quad \text{or} \quad x = 13, \text{ that is, } \bar{x} = 13$$

(b) To find correlation coefficient r between x and y , we need to determine the regression coefficients b_{xy} and b_{yx} .

Rewriting the given regression equations in such a way that the coefficient of dependent variable is less than one at least in one equation.

$$8x - 10y = -66 \quad \text{or} \quad 10y = 66 + 8x \quad \text{or} \quad y = \frac{66}{10} + \frac{8}{10}x$$

That is, $b_{yx} = 8/10 = 0.80$

$$40x - 18y = 214 \quad \text{or} \quad 40x = 214 + 18y \quad \text{or} \quad x = \frac{214}{40} + \frac{18}{40}y$$

That is, $b_{xy} = 18/40 = 0.45$

Hence coefficient of correlation r between x and y is given by

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{0.45 \times 0.80} = 0.60$$

(c) To determine the standard deviation of y , consider the formula:

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} \quad \text{or} \quad \sigma_y = \frac{b_{yx} \sigma_x}{r} = \frac{0.80 \times 3}{0.6} = 4$$

Example 11.9: There are two series of index numbers, P for price index and S for stock of a commodity. The mean and standard deviation of P are 100 and 8 and of S are 103 and 4 respectively. The correlation coefficient between the two series is 0.4. With these

data, work out a linear equation to read off values of P for various values of S. Can the same equation be used to read off values of S for various values of P?

Solution: The regression equation to read off values of P for various values S is given by

$$P = a + bS \quad \text{or} \quad (P - \bar{P}) = r \frac{\sigma_p}{\sigma_s} (S - \bar{S})$$

Given $\bar{P} = 100$, $\bar{S} = 103$, $\sigma_p = 8$, $\sigma_s = 4$, $r = 0.4$. Substituting these values in the above equation, we have

$$P - 100 = 0.4 \frac{8}{4} (S - 103) \quad \text{or} \quad P = 17.6 + 0.8S$$

This equation cannot be used to read off values of S for various values of P. Thus to read off values of S for various values of P we use another regression equation of the form:

$$S = c + dP \quad \text{or} \quad S - \bar{S} = \frac{\sigma_s}{\sigma_p} (P - \bar{P})$$

Substituting given values in this equation, we have

$$S - 103 = 0.4 \frac{4}{8} (P - 100) \quad \text{or} \quad S = 83 + 0.2P$$

Example 11.10: The two regression lines obtained in a correlation analysis of 60 observations are:

$$5x = 6x + 24 \quad \text{and} \quad 1000y = 768x - 3708$$

What is the correlation coefficient and what is its probable error? Show that the ratio of the coefficient of variability of x to that of y is $5/24$. What is the ratio of variances of x and y?

Solution: Rewriting the regression equations

$$5x = 6y + 24 \quad \text{or} \quad x = \frac{6}{5}y + \frac{24}{5}$$

That is, $b_{xy} = 6/5$

$$1000y = 768x - 3708 \quad \text{or} \quad y = \frac{768}{1000}x - \frac{3708}{1000}$$

That is, $b_{yx} = 768/1000$

We know that $b_{xy} = r \frac{\sigma_x}{\sigma_y} = \frac{6}{5}$ and $b_{yx} = r \frac{\sigma_y}{\sigma_x} = \frac{768}{1000}$, therefore

$$b_{xy} b_{yx} = r^2 = \frac{6}{5} \times \frac{768}{1000} = 0.9216$$

Hence $r = \sqrt{0.9216} = 0.96$.

Since both b_{xy} and b_{yx} are positive, the correlation coefficient is positive and hence $r = 0.96$.

$$\begin{aligned} \text{Probable error of } r &= 0.6745 \frac{1-r^2}{\sqrt{n}} = 0.6745 \frac{1-(0.96)^2}{\sqrt{60}} \\ &= \frac{0.0528}{7.7459} = 0.0068 \end{aligned}$$

Solving the given regression equations for x and y, we get $\bar{x} = 6$ and $\bar{y} = 1$ because regression lines passed through the point (\bar{x}, \bar{y}) .

$$\text{Since } r \frac{\sigma_x}{\sigma_y} = \frac{6}{5} \text{ or } 0.96 \frac{\sigma_x}{\sigma_y} = \frac{6}{5} \text{ or } \frac{\sigma_x}{\sigma_y} = \frac{6}{5 \times 0.96} = \frac{5}{4}$$

$$\text{Also the ratio of the coefficient of variability} = \frac{\sigma_x/\bar{x}}{\sigma_y/\bar{y}} = \frac{\bar{y}}{\bar{x}} \cdot \frac{\sigma_x}{\sigma_y} = \frac{1}{6} \times \frac{5}{4} = \frac{5}{24}.$$

11.7.3 Regression Coefficients for Grouped Sample Data

The method of finding the regression coefficients b_{xy} and b_{yx} would be little different than the method discussed earlier for the case when data set is grouped or classified into frequency distribution of either variable x or y or both. The values of b_{xy} and b_{yx} shall be calculated using the formulae:

$$b_{xy} = \frac{n \sum d_x d_y - \sum f d_x \sum f d_y}{n \sum f d_y^2 - (\sum f d_y)^2} \times \frac{h}{k}$$

$$b_{yx} = \frac{n \sum f d_x d_y - \sum f d_x \sum f d_y}{n \sum f d_x^2 - (\sum f d_x)^2} \times \frac{k}{h}$$

where h = width of the class interval of sample data on x variable
 k = width of the class interval of sample data on y variable

Example 11.11: The following bivariate frequency distribution relates to sales turnover (Rs in lakh) and money spent on advertising (Rs in 1000's). Obtain the two regression equations

Sales Turnover (Rs in lakh)	Advertising Budget (Rs in 1000's)			
	50–60	60–70	70–80	80–90
20–50	2	1	2	5
50–80	3	4	7	6
80–110	1	5	8	6
110–140	2	7	9	2

Estimate (a) the sales turnover corresponding to advertising budget of Rs 1,50,000, and (b) the advertising budget to achieve a sales turnover of Rs 200 lakh.

Solution: Let x and y represent sales turnover and advertising budget respectively. Then the regression equation for estimating the sales turnover (x) on advertising budget (y) is expressed as:

$$x - \bar{x} = b_{xy} (y - \bar{y})$$

$$\text{where } b_{xy} = \frac{n \sum f d_x d_y - \sum f d_x \sum f d_y}{n \sum f d_y^2 - (\sum f d_y)^2}$$

Table 11.6: Calculations for Regression Coefficients

		Advertising Budget				f	fd_x	fd_x^2	$fd_x d_y$
		50–60	60–70	70–80	80–90				
$Sales$	x	y	$m.v$	d_y	f				
20–50	35	1	2	(4)	1	(1)	2	—	5 (−5)
50–80	65	0	3	—	4	—	7	—	6 (—)
80–110	95	1	1	(−2)	5	(−5)	8	—	6 (6)
110–140	125	2	2	(−8)	7	(−14)	9	—	2 (4)
		f	8	17	26	19	$n = 70$	$50 = \Sigma f d_x$	$110 = \Sigma f d_x^2$
		fd_y	−16	−17	0	19		$-14 = \Sigma f d_y$	$-19 = \Sigma f d_x d_y$
		fd_y^2	32	17	0	19		$68 = \Sigma f d_y^2$	
		$fd_x d_y$	−6	−18	0	5		$-19 = \Sigma f d_x d_y$	

Similarly, the regression equation for estimating the advertising budget (y) on sales turnover of Rs 200 lakh is written as:

$$y - \bar{y} = b_{yx} (x - \bar{x})$$

where $b_{yx} = \frac{n \sum f d_x d_y - (\sum f d_x)(\sum f d_y)}{n \sum f d_x^2 - (\sum f d_x)^2}$

The calculations for regression coefficients b_{xy} and b_{yx} are shown in Table 11.6.

$$\bar{x} = A + \frac{\sum f d_x}{n} \times h = 65 + \frac{50}{70} \times 30 = 65 + 21.428 = 86.428$$

$$\bar{y} = B + \frac{\sum f d_y}{n} \times k = 75 - \frac{14}{70} \times 10 = 75 - 2 = 73$$

$$b_{xy} = \frac{n \sum f d_x d_y - (\sum f d_x)(\sum f d_y)}{n \sum f d_y^2 - (\sum f d_y)^2} \times \frac{h}{k} = \frac{70 \times -19 - (50)(-14)}{70 \times 68 - (-14)^2} \times \frac{30}{10}$$

$$= \frac{-1330 + 700}{4760 - 196} \times \frac{30}{10} = \frac{-6300}{45,640} = -0.414$$

$$b_{yx} = \frac{n \sum f d_x d_y - (\sum f d_x)(\sum f d_y)}{n \sum f d_x^2 - (\sum f d_x)^2} \times \frac{k}{h} = \frac{70 \times -19 - (50)(-14)}{70 \times 110 - (50)^2} \times \frac{10}{30}$$

$$= \frac{-1330 + 700}{7700 - 2500} \times \frac{10}{30} = \frac{-6300}{1,56,000} = -0.040$$

Substituting these values in the two regression equations, we get

(a) Regression equation of sales turnover (x) to advertising budget (y) is:

$$x - \bar{x} = b_{xy} (y - \bar{y})$$

$$x - 86.428 = -0.414 (y - 73), \text{ or } x = 116.65 - 0.414y$$

For $y = 150$, we have $x = 116.65 - 0.414 \times 150 = \text{Rs } 54.55 \text{ lakh}$

(b) Regression equation of advertising budget (y) on sales turnover (x) is:

$$y - \bar{y} = b_{yx} (x - \bar{x})$$

$$y - 73 = -0.040 (x - 86.428) \text{ or } y = 76.457 - 0.04x$$

For $x = 200$, we have $y = 76.457 - 0.04 (200) = \text{Rs } 68.457 \text{ thousand.}$

Self-Practice Problems 11A

- 11.1** The following calculations have been made for prices of twelve stocks (x) at the Calcutta Stock Exchange on a certain day along with the volume of sales in thousands of shares (y). From these calculations find the regression equation of price of stocks on the volume of sales of shares.

$$\Sigma x = 580, \quad \Sigma y = 370, \quad \Sigma xy = 11494,$$

$$\Sigma x^2 = 41658, \quad \Sigma y^2 = 17206.$$

[Rajasthan Univ., MCom, 1995]

- 11.2** A survey was conducted to study the relationship between expenditure (in Rs) on accommodation (x) and expenditure on food and entertainment (y) and the following results were obtained:

	Mean	Standard Deviation
• Expenditure on accommodation	173	63.15
• Expenditure on food and entertainment	47.8	22.98
Coefficient of correlation $r = 0.57$		

Write down the regression equation and estimate the expenditure on food and entertainment if the expenditure on accommodation is Rs 200.

[Bangalore Univ., BCom, 1998]

- 11.3** The following data give the experience of machine operators and their performance ratings given by the number of good parts turned out per 100 pieces:

Operator : 1 2 3 4 5 6 7 8
experience (x) : 16 12 18 4 3 10 5 12

Performance ratings (y) : 87 88 89 68 78 80 75 83
Calculate the regression lines of performance ratings on experience and estimate the probable performance if an operator has 7 years experience.

[Jammu Univ., MCom; Lucknow Univ., MBA, 1996]

- 11.4** A study of prices of a certain commodity at Delhi and Mumbai yield the following data:

	<i>Delhi</i>	<i>Mumbai</i>
• Average price per kilo (Rs)	2.463	2.797
• Standard deviation	0.326	0.207
• Correlation coefficient between prices at Delhi and Mumbai $r = 0.774$		

Estimate from the above data the most likely price (a) at Delhi corresponding to the price of Rs 2.334 per kilo at Mumbai (b) at Mumbai corresponding to the price of 3.052 per kilo at Delhi.

- 11.5 The following table gives the aptitude test scores and productivity indices of 10 workers selected at random:

Aptitude scores (x) : 60 62 65 70 72 48 53 73 65 82
 Productivity index (y) : 68 60 62 80 85 40 52 62 60 81

Calculate the two regression equations and estimate (a) the productivity index of a worker whose test score is 92, (b) the test score of a worker whose productivity index is 75. [Delhi Univ., MBA, 2001]

- 11.6 A company wants to assess the impact of R&D expenditure (Rs in 1000s) on its annual profit; (Rs in 1000's). The following table presents the information for the last eight years:

<i>Year</i>	<i>R & D expenditure</i>	<i>Annual profit</i>
1991	9	45
1992	7	42
1993	5	41
1994	10	60
1995	4	30
1996	5	34
1997	3	25
1998	2	20

Estimate the regression equation and predict the annual profit for the year 2002 for an allocated sum of Rs 1,00,000 as R&D expenditure.

[Jodhpur Univ., MBA, 1998]

- 11.7 Obtain the two regression equations from the following bivariate frequency distribution:

<i>Sales Revenue</i> (Rs in lakh)	<i>Advertising Expenditure (Rs in thousand)</i>			
	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	3	3	2	2

Estimate (a) the sales corresponding to advertising expenditure of Rs 50,000, (b) the advertising expenditure for a sales revenue of Rs 300 lakh, (c) the coefficient of correlation. [Delhi Univ., MBA, 2002]

- 11.8 The personnel manager of an electronic manufacturing company devises a manual test for job applicants to predict their production rating in the assembly

department. In order 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	A	B	C	D	E	F	G	H	I	J
Test score	53	36	88	84	86	64	45	48	39	69
Production rating	45	43	89	79	84	66	49	48	43	76

Fit a linear least squares regression equation of production rating on test score. [Delhi Univ., MBA, 200]

- 11.9 Find the regression equation showing the capacity utilization on production from the following data:

	<i>Average Deviation</i>	<i>Standard Deviation</i>
• Production (in lakh units)	35.6	10.5
• Capacity utilization (in percentage)	84.8	8.5
• Correlation coefficient $r = 0.62$		

Estimate the production when the capacity utilization is 70 per cent.

[Delhi Univ., MBA, 1997; Pune Univ., MBA, 1998]

- 11.10 Suppose that you are interested in using past expenditure on R&D by a firm to predict current expenditures on R&D. You got the following data by taking a random sample of firms, where x is the amount spent on R&D (in lakh of rupees) 5 years ago and y is the amount spent on R&D (in lakh of rupees) in the current year:

x : 30 50 20 80 10 20 20 40
 y : 50 80 30 110 20 20 40 50

- (a) Find the regression equation of y on x .
 (b) If a firm is chosen randomly and $x = 10$, can you use the regression to predict the value of y ? Discuss.

[Madurai-Kamraj Univ., MBA, 2000]

- 11.11 The following data relates to the scores obtained by a salesmen of a company in an intelligence test and their weekly sales (in Rs. 1000's):

Salesman	intelligence	A	B	C	D	E	F	G	H	I
Test score	: 50 60 50 60 80 50 80 40 70									
Weekly sales	: 30 60 40 50 60 30 70 50 60									

- (a) Obtain the regression equation of sales on intelligence test scores of the salesmen.
 (b) If the intelligence test score of a salesman is 65, what would be his expected weekly sales?

[HP Univ., M.com., 1996]

- 11.12 Two random variables have the regression equations:

$$3x + 2y - 26 = 0 \quad \text{and} \quad 6x + y - 31 = 0$$

- (a) Find the mean values of x and y and coefficient of correlation between x and y .
 (b) If the variance of x is 25, then find the standard deviation of y from the data.

[MD Univ., M.Com., 1997; Kumaun Univ., MBA, 2001]

- 11.13** For a given set of bivariate data, the following results were obtained

$$\bar{x} = 53.2, \bar{y} = 27.9,$$

Regression coefficient of y on $x = -1.5$, and Regression coefficient of x and $y = -0.2$.

Find the most probable value of y when $x = 60$.

- 11.14** In trying to evaluate the effectiveness in its advertising campaign, a firm compiled the following information: Calculate the regression equation of sales on advertising expenditure. Estimate the probable sales when advertisement expenditure is Rs. 60 thousand.

Year	Adv. expenditure (Rs. 1000's)	Sales (in lakhs Rs)
1996	12	5.0
1997	15	5.6
1998	17	5.8
1999	23	7.0
2000	24	7.2
2001	38	8.8
2002	42	9.2
2003	48	9.5

[Bharathidasan Univ., MBA, 2003]

Hints and Answers

11.1 $\bar{x} = \Sigma x/n = 580/12 = 48.33$;

$$\bar{y} = \Sigma y/n = 370/12 = 30.83$$

$$b_{xy} = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma y^2 - n(\bar{y})^2} = \frac{11494 - 12 \times 48.33 \times 30.83}{17206 - 12(30.83)^2} = -1.102$$

Regression equation of x on y :

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 48.33 = -1.102(y - 30.83)$$

$$\text{or } x = 82.304 - 1.102y$$

- 11.2** Given $\bar{x} = 172$, $\bar{y} = 47.8$, $\sigma_x = 63.15$, $\sigma_y = 22.98$, and $r = 0.57$

Regression equation of food and entertainment (y) on accomodation (x) is given by

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$y - 47.8 = 0.57 \frac{22.98}{63.15} (x - 172)$$

$$\text{or } y = 11.917 + 0.207x$$

For $x = 200$, we have $y = 11.917 + 0.207(200) = 53.317$

- 11.3** Let the experience and performance rating be represented by x and y respectively.

$$\bar{x} = \Sigma x/n = 80/8 = 10; \bar{y} = \Sigma y/n = 648/8 = 81$$

$$b_{yx} = \frac{n \sum d_x d_y - \sum d_x \sum d_y}{n \sum d_x^2 - (\sum d_x)^2} = \frac{247}{218} = 1.133;$$

where $d_x = x - \bar{x}$, $d_y = y - \bar{y}$

Regression equation of y on x

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\text{or } y - 81 = 1.133(x - 10)$$

$$\text{or } y = 69.67 + 1.133x$$

When $x = 7$, $y = 69.67 + 1.133(7) = 77.60 \cong 78$

- 11.4** Let price at Mumbai and Delhi be represented by x and y , respectively

(a) Regression equation of y on x

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$y - 2.463 = 0.774 \frac{0.326}{0.207} (x - 2.797)$$

For $x = \text{Rs } 2.334$, the price at Delhi would be $y = \text{Rs } 1.899$.

(b) Regression equation of x on y

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$\text{or } x - 2.791 = 0.774 \frac{0.207}{0.326} (y - 2.463)$$

For $y = \text{Rs } 3.052$, the price at Mumbai would be $x = \text{Rs } 3.086$.

- 11.5** Let aptitude score and productivity index be represented by x and y respectively.

$$\bar{x} = \Sigma x/n = 650/10 = 65; \bar{y} = \Sigma y/n = 650/10 = 65$$

$$b_{xy} = \frac{n \sum d_x d_y - (\sum d_x)(\sum d_y)}{n \sum d_y^2 - (\sum d_y)^2} = \frac{1044}{1752} = 0.596;$$

where $d_x = x - \bar{x}$; $d_y = y - \bar{y}$

(a) Regression equation of x on y

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$\text{or } x - 65 = 0.596(y - 65)$$

$$\text{or } x = 26.26 + 0.596y$$

When $y = 75$, $x = 26.26 + 0.596(75) = 70.96 \cong 71$

$$(b) b_{yx} = \frac{n \sum d_x d_y - (\sum d_x)(\sum d_y)}{n \sum d_x^2 - (\sum d_x)^2} = \frac{1044}{894} = 1.168$$

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\text{or } y - 65 = 1.168(x - 65)$$

$$\text{or } y = -10.92 + 1.168x$$

When $x = 92$, $y = -10.92 + 1.168(92) = 96.536 \cong 97$

- 11.6** Let R&D expenditure and annual profit be denoted by x and y respectively

$$\bar{x} = \Sigma x/n = 40/8 = 5.625; \bar{y} = \Sigma y/n = 297/8 = 37.125$$

$$b_{yx} = \frac{n \Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n \Sigma d_x^2 - (\Sigma d_x)^2} = \frac{8 \times 238 - (-3)(1)}{8 \times 57 - (-3)^2} = 4.266;$$

where $d_x = x - 6$, $d_y = y - 37$

Regression equation of annual profit on R&D expenditure

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 37.125 = 4.26(x - 5.625)$$

or $y = 13.163 + 4.266x$

For $x = \text{Rs } 1,00,000$ as R&D expenditure, we have from above equation $y = \text{Rs } 439.763$ as annual profit.

- 11.7** Let sales revenue and advertising expenditure be denoted by x and y respectively

$$\bar{x} = A + \frac{\Sigma f d_x}{n} \times h = 150 + \frac{12}{66} \times 50 = 159.09$$

$$\bar{y} = B + \frac{\Sigma f d_y}{n} \times k = 30 - \frac{26}{66} \times 10 = 26.06$$

$$b_{xy} = \frac{n \Sigma f d_x d_y - (\Sigma f d_x)(\Sigma f d_y)}{n \Sigma f d_y^2 - (\Sigma f d_y)^2} \times \frac{h}{k}$$

$$= \frac{66(-14) - 12(-26)}{66(100) - (-26)^2} \times \frac{50}{10} = -0.516$$

(a) Regression equation of x on y

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 159.09 = -0.516(y - 26.06)$$

or $x = 172.536 - 0.516y$

For $y = 50$, $x = 147.036$

(b) Regression equation of y on x

$$b_{yx} = \frac{n \Sigma f d_x d_y - (\Sigma f d_x)(\Sigma f d_y)}{n \Sigma f d_x^2 - (\Sigma f d_x)^2} \times \frac{k}{h}$$

$$= \frac{66(-14) - 12(-26)}{66(70) - (12)^2} \times \frac{10}{50} = -0.027.$$

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 26.06 = -0.027(x - 159.09)$$

$$y = 30.355 - 0.027x$$

For $x = 300$, $y = 22.255$

$$(c) r = \sqrt{b_{xy} \times b_{yx}} = -\sqrt{0.516 \times 0.027} = -0.1180$$

- 11.8** Let test score and production rating be denoted by x and y respectively.

$$\bar{x} = \Sigma x/n = 612/10 = 61.2;$$

$$\bar{y} = \Sigma y/n = 622/10 = 62.2$$

$$b_{yx} = \frac{n \Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n \Sigma d_x^2 - (\Sigma d_x)^2} = \frac{10 \times 3213 - 2 \times 2}{10 \times 3554 - (2)^2} = 0.904$$

Regression equation of production rating (y) on test score (x) is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 62.2 = 0.904(x - 61.2)$$

$$y = 6.876 + 0.904x$$

- 11.9** Let production and capacity utilization be denoted by x and y , respectively.

- (a) Regression equation of capacity utilization (y) on production (x)

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$y - 84.8 = 0.62 \frac{8.5}{10.5} (x - 35.6)$$

$$y = 66.9324 + 0.5019x$$

- (b) Regression equation of production (x) on capacity utilization (y)

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$x - 35.6 = 0.62 \frac{10.5}{8.5} (y - 84.8)$$

$$x = -29.3483 + 0.7659y$$

When $y = 70$, $x = -29.3483 + 0.7659(70) = 24.2647$

Hence the estimated production is 2,42,647 units when the capacity utilization is 70 per cent.

- 11.10** $\bar{x} = \Sigma x/n = 270/8 = 33.75$; $\bar{y} = \Sigma y/n = 400/8 = 50$

$$b_{yx} = \frac{n \Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n \Sigma d_x^2 - (\Sigma d_x)^2} = \frac{8 \times 4800 - 6 \times 0}{8 \times 3592 - (6)^2} = 1.338;$$

where $d_x = x - 33$ and $d_y = y - 50$

Regression equation of y on x

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 50 = 1.338(x - 33.75)$$

$$y = 4.84 + 1.338x$$

For $x = 10$, $y = 18.22$

- 11.11** Let intelligence test score be denoted by x and weekly sales by y

$$\bar{x} = 540/9 = 60; \bar{y} = 450/9 = 50,$$

$$b_{yx} = \frac{n \Sigma dx dy - (\Sigma dx)(\Sigma dy)}{n \Sigma d_x^2 - (\Sigma d_x)^2} = \frac{9 \times 1200}{9 \times 1600} = 0.75$$

Regression equation of y on x :

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 50 = 0.75(x - 60)$$

$$y = 5 + 0.75x$$

For $x = 65$, $y = 5 + 0.75(65) = 53.75$

- 11.12** (a) Solving two regression lines:

$$3x + 2y = 6 \quad \text{and} \quad 6x + y = 31$$

we get mean values as $\bar{x} = 4$ and $\bar{y} = 7$

(b) Rewriting regression lines as follows:

$$3x + 2y = 6 \quad \text{or} \quad y = 13 - (3/2)x,$$

$$\text{So } b_{yx} = -3/2$$

$$6x + y = 31 \quad \text{or} \quad x = 31/6 - (1/6)y,$$

$$\text{So } b_{xy} = -1/6$$

Correlation coefficient,

$$r = \sqrt{b_{xy} \times b_{yx}} = -\sqrt{(3/2)(1/6)} = -0.5$$

Given, $\text{Var}(x) = 25$, so $\sigma_x = 5$. Calculate σ_y using the formula:

$$b_{yx} = r \frac{\sigma_y}{\sigma_x}$$

$$\text{or } -\frac{3}{2} = 0.5 \frac{\sigma_y}{5} \quad \text{or } \sigma_y = 15$$

11.13 The regression equation of y on x is stated as:

$$y - \bar{y} = b_{yx}(x - \bar{x}) = r \cdot \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

Given, $\bar{x} = 53.20$; $\bar{y} = 27.90$, $b_{yx} = -1.5$; $b_{xy} = -0.2$

Thus $y - 27.90 = -1.5(x - 53.20)$

$$\text{or } y = 107.70 - 1.5x$$

For $x = 60$, we have $y = 107.70 - 1.5(60) = 17.7$

$$\text{Also } r = \sqrt{b_{yx} \times b_{xy}} = -\sqrt{1.5 \times 0.2} = -0.5477$$

11.14 Let advertising expenditure and sales be denoted by x and y respectively.

$$\bar{x} = \Sigma x/n = 217/8 = 27.125; \bar{y} = \Sigma y/n = 58.2/8 = 7.26$$

$$\begin{aligned} b_{yx} &= \frac{n \sum dx dy - (\sum dx)(\sum dy)}{n \sum d_x^2 - (\sum dx)^2} \\ &= \frac{8(172.2) - (25)(2.1)}{8(1403) - (25)^2} = \frac{1325.1}{10599} = 0.125 \end{aligned}$$

Thus regression equation of y on x is:

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\text{or } y - 7.26 = 0.125(x - 27.125)$$

$$y = 3.86 + 0.125x$$

When $x = 60$, the estimated value of $y = 3.869 + 0.125(60) = 11.369$

11.8 STANDARD ERROR OF ESTIMATE AND PREDICTION INTERVALS

The pattern of dot points on a scatter diagram is an indicator of the relationship between two variables x and y . Wide scatter or variation of the dot points about the regression line represents a poor relationship. But a very close scatter of dot points about the regression line represents a close relationship between two variables. The variability in observed values of dependent variable y about the regression line is measured in terms of *residuals*. A residual is defined as the difference between an observed value of dependent variable y and its estimated (or fitted) value \hat{y} determined by regression equation for a given value of the independent variable x . The residual about the regression line is given by

$$\text{Residual } e_i = y_i - \hat{y}_i$$

The residual values e_i are plotted on a diagram with respect to the least squares regression line $\hat{y} = a + bx$. These residual values represent error of estimation for individual values of dependent variable and are used to estimate, the variance σ^2 of the error term. In other words, residuals are used to estimate the amount of variation in the dependent variable with respect to least squares regression line. Here it should be noted that the variations are not the variations (deviations) of observations from the mean value in the sample data set, rather these variations are the vertical distances of every observation (dot point) from the least squares line as shown in Fig. 11.3.

Since sum of the residuals is zero, therefore it is not possible to determine the total amount of error by summing the residuals. This zero-sum characteristic of residuals can be avoided by squaring the residuals and then summing them. That is

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \leftarrow \text{Error or Residual sum of squares}$$

This quantity is called the *sum of squares of errors (SSE)*.

The estimate of variance of the error term σ_e^2 or $S_{y,x}^2$ is obtained as follows:

$$S_{y,x}^2 \text{ or } \hat{\sigma}_e^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$$

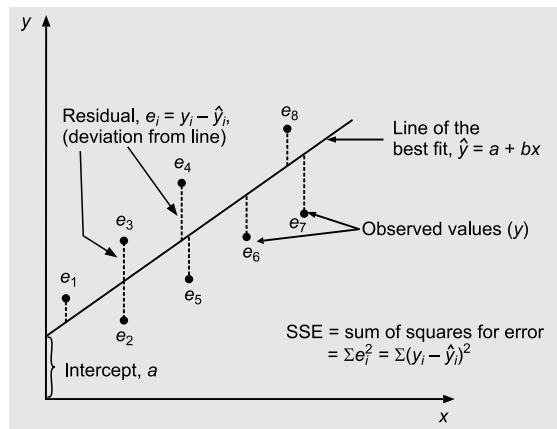
The denominator, $n - 2$ represents the *error or residual degrees of freedom* and is determined by subtracting from sample size n the number of parameters β_0 and β_1 that are estimated by the sample parameters a and b in the least squares equation. The subscript 'yx' indicates that the standard deviation is of dependent variable y , given (or conditional) upon independent variable x .

The *standard error of estimate* $S_{y,x}$ also called *standard deviation of the error term* measures the variability of the observed values around the regression line, i.e. the amount

by which the \hat{y} values are away from the sample y values (dot points). In other words, S_{yx} is based on the deviations of the sample observations of y -values from the least squares line or the estimated regression line of \hat{y} values. The standard deviation of error about the least squares line is defined as:

$$S_{yx} \text{ or } \sigma_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}} = \sqrt{\frac{\text{SSE}}{n - 2}} \quad (11-4)$$

Figure 11.3
Residuals



To simplify the calculations of S_{yx} , generally the following formula is used

$$S_{yx} = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}} = \sqrt{\frac{\sum y^2 - a \sum y - b \sum xy}{n - 2}}$$

The variance S_{yx}^2 measures how the least squares line 'best fits' the sample y -values. A large variance and standard error of estimate indicates a large amount of scatter or dispersion of dot points around the line. Smaller the value of S_{yx} , the closer the dot points (y -values) fall around the regression line and better the line fits the data and describes the better average relationship between the two variables. When all dot points fall on the line, the value of S_{yx} is zero, and the relationship between the two variables is perfect.

A smaller variance about the regression line is considered useful in predicting the value of a dependent variable y . In actual practice, some variability is always left over about the regression line. It is important to measure such variability due to the following reasons:

- (i) This value provides a way to determine the usefulness of the regression line in predicting the value of the dependent variable.
- (ii) This value can be used to construct interval estimates of the dependent variable.
- (iii) Statistical inferences can be made about other components of the problem.

Figure 11.4 displays the distribution of conditional average values of y about a least squares regression line for given values of independent variable x . Suppose the amount of deviation in the values of y given any particular value of x follow normal distribution. Since average value of y changes with the value of x , we have different normal distributions of y -values for every value of x , each having same standard deviation. When a relationship between two variables x and y exists, the standard deviation (also called *standard error of estimate*) is less than the standard deviation of all the x -values in the population computed about their mean.

Based on the assumptions of regression analysis, we can describe sampling properties of the sample estimates such as a , b , and S_{yx} , as these vary from sample to sample. Such knowledge is useful in making statistical inferences about the relationship between the two variables x and y .

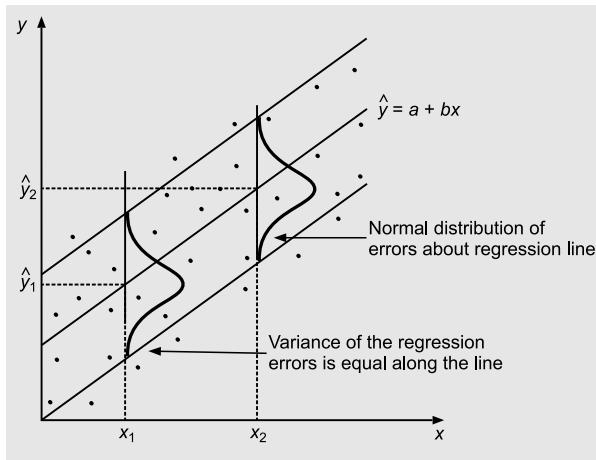


Figure 11.4
Regression Line Showing the Error Variance

The standard error of estimate can also be used to determine an approximate interval estimate based on sample data ($n < 30$) for the value of the dependent variable y for a given value of the independent variable x as follows:

$$\text{Approximate interval estimate} = \hat{y} \pm t_{df} S_{yx}$$

where value of t is obtained using t -distribution table based upon a chosen probability level. The interval estimate is also called a *prediction interval*.

Example 11.12: The following data relate to advertising expenditure (Rs in lakh) and their corresponding sales (Rs in crore)

Advertising expenditure : 10 12 15 23 20

Sales : 14 17 23 25 21

- Find the equation of the least squares line fitting the data.
- Estimate the value of sales corresponding to advertising expenditure of Rs 30 lakh.
- Calculate the standard error of estimate of sales on advertising expenditure.

Solution: Let the advertising expenditure be denoted by x and sales by y .

(a) The calculations for the least squares line are shown in Table 11.7

Table 11.7: Calculations for Least Squares Line

Advt. Expenditure, x	$d_x = x - 16$	d_x^2	Sales y	$d_y = y - 20$	d_y^2	$d_x d_y$
10	-6	36	14	-6	36	36
12	-4	16	17	-3	9	12
15	-1	1	23	3	9	-3
23	7	49	25	5	25	35
20	4	16	21	1	1	4
80	0	118	100	0	80	84

$$\bar{x} = \Sigma x/n = 80/5 = 16; \quad \bar{y} = \Sigma y/n = 100/5 = 20$$

$$b_{yx} = \frac{n \sum d_x d_y - \sum d_x \sum d_y}{n \sum d_x^2 - (\sum d_x)^2} = \frac{5 \times 84}{5 \times 118} = 0.712$$

(a) Regression equation of y on x is

$$y - \bar{y} = b_{yx} (x - \bar{x})$$

$$y - 20 = 0.712 (x - 16)$$

$$y = 8.608 + 0.712 x$$

where parameter $a = 8.608$ and $b = 0.712$.

Table 11.8 gives the fitted values and the residuals for the data in Table 11.7. The fitted values are obtained by substituting the value of x into the regression equation (equation for the least squares line). For example, $8.608 + 0.712(10) = 15.728$. The

residual is equal to the actual value minus fitted value. The residuals indicate how well the least squares line fits the actual data values.

Table 11.8: Fitted Values and Residuals for Sample Data

Value, x	Fitted Value	Residuals
	$y = 8.608 + 0.712x$	
10	15.728	-5.728
12	17.152	-5.152
15	19.288	-4.288
23	24.984	-1.984
20	22.848	-2.848

(b) The least squares equation obtained in part (a) may be used to estimate the sales turnover corresponding to the advertising expenditure of Rs 30 lakh as:

$$\hat{y} = 8.608 + 0.712x = 8.608 + 0.712(30) = \text{Rs } 29.968 \text{ crore}$$

(c) Calculations for standard error of estimate $S_{y|x}$ of sales (y) on advertising expenditure (x) are shown in Table 11.9.

Table 11.9: Calculations for Standard Error of Estimate

x	y	y^2	xy
10	14	196	140
12	17	289	204
15	23	529	345
23	25	625	575
20	21	441	420
80	100	2080	1684

$$S_{y|x} = \sqrt{\frac{\sum y^2 - a \sum y - b \sum xy}{n-2}} = \sqrt{\frac{2080 - 8.608 \times 100 - 0.712 \times 1684}{5-2}}$$

$$= \sqrt{\frac{2080 - 860.8 - 1199}{3}} = 2.594$$

11.8.1 Coefficient of Determination: Partitioning of Total Variation

The objective of regression analysis is to develop a regression model that best fits the sample data, so that the residual variance $S_{y|x}^2$ is as small as possible. But the value of $S_{y|x}^2$ depends on the scale with which the sample y -values are measured. This drawback with the calculation of $S_{y|x}^2$ restricts its interpretation unless we consider the units in which the y -values are measured. Thus, we need another measure of fit called *coefficient of determination* that is not affected by the scale with which the sample y -values are measured. It is the proportion of variability of the dependent variable, y accounted for or explained by the independent variable, x , i.e. it measures how well (i.e. strength) the regression line fits the data. The coefficient of determination is denoted by r^2 and its value ranges from 0 to 1. A particular r^2 value should be interpreted as high or low depending upon the use and context in which the regression model was developed. The coefficient of determination is given by

$$r^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

$$= 1 - \frac{\text{Residual variation in response variable } y\text{-values from least-squares line}}{\text{Total variance of } y\text{-values}}$$

where $SST =$ total sum of square deviations (or total variance) of sampled response variable y -values from the mean value of y .

$$= S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n(\bar{y})^2$$

$SSE =$ sum of squares of error or *unexplained variation* in response variable y -values from the least squares line due to sampling errors, i.e. it measures the residual variation in the data that is not explained by predictor variable x

$$= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n y_i^2 - a \sum_{i=1}^n y_i - b \sum_{i=1}^n x_i y_i$$

$SSR =$ sum of squares of regression or *explained variation* is the sample values of response variable y accounted for or explained by variation among x -values

$$= SST - SSE$$

$$= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = a \sum_{i=1}^n y_i + b \sum_{i=1}^n x_i y_i - n(\bar{y})^2$$

The three variations associated with the regression analysis of a data set are shown in Fig 11.5. Thus

$$r^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2} = 1 - \frac{S_{yx}^2}{S_y^2}; \quad S_{yx} = S_y \sqrt{1 - r^2}$$

where $\frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$ = fraction of the total variation that is explained or accounted for

$S_{y.x} = \frac{\Sigma(y - \hat{y})^2}{n - 2}$, variance of response variable y -values from the least squares line

$$S_y^2 = \frac{1}{n - 2} \Sigma(y - \bar{y})^2, \text{ total variance of response variable } y\text{-values}$$

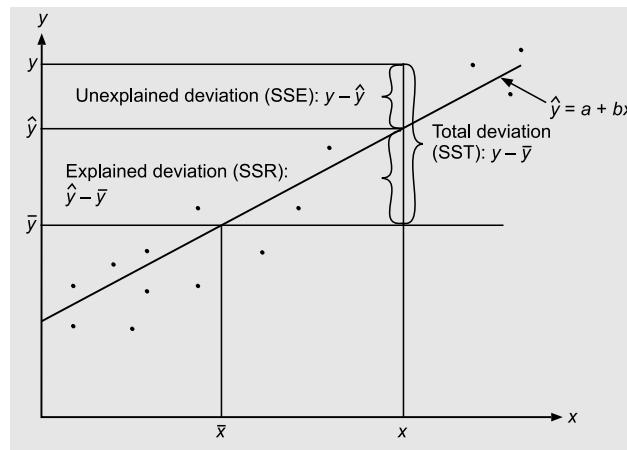


Figure 11.5
Relationship Between Three Types of Variations

Since the formula of r^2 is not convenient to use therefore an easy formula for the sample coefficient of determination is given by

$$r^2 = \frac{a \Sigma y + b \Sigma xy - n(\bar{y})^2}{\Sigma y^2 - n(\bar{y})^2} \leftarrow \text{Short-cut method}$$

For example, the coefficient of determination that indicates the extent of relationship between sales revenue (y) and advertising expenditure (x) is calculated as follows from Example 11.1:

$$\begin{aligned}
 r^2 &= \frac{a \sum y + b \sum xy - n(\bar{y})^2}{\sum y^2 - n(\bar{y})^2} = \frac{0.072 \times 40 + 0.704 \times 373 - 8(5)^2}{270 - 8(5)^2} \\
 &= \frac{2.88 + 262.592 - 200}{270 - 200} = \frac{65.47}{70} = 0.9352
 \end{aligned}$$

The value $r^2 = 0.9352$ indicates that 93.52% of the variance in sales revenue is accounted for or statistically explained by advertising expenditure.

A comparison between bivariate correlation and regression summarized in Table 11-10 could provide further insight about the relationship between two variables x and y in the data set.

Table 11.10: Comparison between Linear Correlation and Regression

	<i>Correlation</i>	<i>Regression</i>
• Measurement level	Interval or ratio scale	Interval or ratio scale
• Nature of variables	Both continuous, and linearly related	Both continuous, and linearly related
• $x-y$ relationship	x and y are symmetric	y is dependent, x is independent; regression of x on y differs from y on x
• Correlation	$b_{xy} = b_{yx}$	Correlation between x and y is the same as the correlation between y and x
• Coefficient of determination	Explains common variance of x and y	Proportion of variability of x explained by its least-squares regression on y

Conceptual Questions 11A

1. (a) Explain the concept of regression and point out its usefulness in dealing with business problems.

[Delhi Univ., MBA, 1993]

- (b) Distinguish between correlation and regression. Also point out the properties of regression coefficients.

2. Explain the concept of regression and point out its importance in business forecasting.

[Delhi Univ., MBA, 1990, 1998]

3. Under what conditions can there be one regression line? Explain.

[HP Univ., MBA, 1996]

4. Why should a residual analysis always be done as part of the development of a regression model?

5. What are the assumptions of simple linear regression analysis and how can they be evaluated?

6. What is the meaning of the standard error of estimate?

7. What is the interpretation of y -intercept and the slope in a regression model?

8. What are regression lines? With the help of an example illustrate how they help in business decision-making.

[Delhi Univ., MBA, 1998]

9. Point out the role of regression analysis in business decision-making. What are the important properties of regression coefficients?

[Osmania Univ., MBA; Delhi Univ., MBA, 1999]

10. (a) Distinguish between correlation and regression analysis.

[Dipl in Mgt., AIMA, Osmania Univ., MBA, 1998]

- (b) The coefficient of correlation and coefficient of determination are available as measures of association in correlation analysis. Describe the different uses of these two measures of association.

11. What are regression coefficients? State some of the important properties of regression coefficients.

[Dipl in Mgt., AIMA, Osmania Univ., MBA, 1989]

12. What is regression? How is this concept useful to business forecasting?

[Jodhpur Univ., MBA, 1999]

13. What is the difference between a prediction interval and a confidence interval in regression analysis?

14. Explain what is required to establish evidence of a cause-and-effect relationship between y and x with regression analysis.

15. What technique is used initially to identify the kind of regression model that may be appropriate.
16. (a) What are regression lines? Why is it necessary to consider two lines of regression?
 (b) In case the two regression lines are identical, prove that the correlation coefficient is either + 1 or - 1. If two variables are independent, show that the two regression lines cut at right angles.
17. What are the purpose and meaning of the error terms in regression?
18. Give examples of business situations where you believe a straight line relationship exists between two variables. What would be the uses of a regression model in each of these situations.
19. 'The regression lines give only the best estimate of the value of quantity in question. We may assess the degree of uncertainty in the estimate by calculating a quantity known as the standard error of estimate' Elucidate.
20. Explain the advantages of the least-squares procedure for fitting lines to data. Explain how the procedure works.

Formulae Used

1. Simple linear regression model

$$y = \beta_0 + \beta_1 x + e$$

2. Simple linear regression equation based on sample data

$$y = a + bx$$

3. Regression coefficient in sample regression equation

$$b = \hat{y}$$

$$a = \bar{y} - b\bar{x}$$

4. Residual representing the difference between an observed value of dependent variable y and its fitted value

$$e = y - \hat{y}$$

5. Standard error of estimate based on sample data

- Deviations formula

$$S_{y,x} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$$

- Computational formula

$$S_{y,x} = \sqrt{\frac{\sum y^2 - a \sum y - b \sum xy}{n - 2}}$$

6. Coefficient of determination based on sample data

- Sums of squares formula

$$r^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

- Computational formula

$$r^2 = \frac{a \sum y + b \sum xy - n(\bar{y})^2}{\sum y^2 - n(\bar{y})^2}$$

7. Regression sum of squares

$$S_{y,x} = S_y \sqrt{1 - r^2}$$

8. Interval estimate based on sample data: $\hat{y} \pm t_{df} S_{y,x}$

Review Self-Practice Problems

- 11.15** Given the following bivariate data:

x:	-1	5	3	2	1	1	7	3
y:	-6	1	0	0	1	2	1	5

- (a) Fit a regression line of y on x and predict y if $x = 10$.
 (b) Fit a regression line of x on y and predict x if $y = 2.5$.

[Osmania Univ., MBA, 1996]

- 11.16** Find the most likely production corresponding to a rainfall of 40 inches from the following data:

	Rainfall (in inches)	Production (in quintals)
Average	30	50
Standard deviation	5	10

Coefficient of correlation $r = 0.8$.

[Bharthidarsan Univ., MCom, 1996]

- 11.17** The coefficient of correlation between the ages of husbands and wives in a community was found to be + 0.8, the average of husbands age was 25 years and that of wives age 22 years. Their standard deviations were 4 and 5 years respectively. Find with the help of regression equations:

- (a) the expected age of husband when wife's age is 16 years, and
 (b) the expected age of wife when husband's age is 33 years.

[Osmania Univ., MBA, 2000]

- 11.18** You are given below the following information about advertisement expenditure and sales:

	Adv. Exp. (x) (Rs in crore)	Sales (y) (Rs in crore)
Mean	20	120
Standard deviation	5	25

Correlation coefficient 0.8

- Calculate the two regression equations.
- Find the likely sales when advertisement expenditure is Rs 25 crore.
- What should be the advertisement budget if the company wants to attain sales target of Rs 150 crore?

[Jammu Univ., MCom, 1997; Delhi Univ., MBA, 1999]

- 11.19** For 50 students of a class the regression equation of marks in Statistics (x) on the marks in Accountancy (y) is $3y - 5x + 180 = 0$. The mean marks in Accountancy is 44 and the variance of marks in Statistics is $9/16$ th of the variance of marks in Accountancy. Find the mean marks in Statistics and the coefficient of correlation between marks in the two subjects.

- 11.20** The HRD manager of a company wants to find a measure which he can use to fix the monthly income of persons applying for a job in the production department. As an experimental project, he collected data on 7 persons from that department referring to years of service and their monthly income.

Years of service : 11 7 9 5 8 6 10
Income (Rs in 1000's) : 10 8 6 5 9 7 11

- Find the regression equation of income on years of service.
- What initial start would you recommend for a person applying for the job after having served in a similar capacity in another company for 13 years?
- Do you think other factors are to be considered (in addition to the years of service) in fixing the income with reference to the above problems? Explain.

- 11.21** The following table gives the age of cars of a certain make and their annual maintenance costs. Obtain the regression equation for costs related to age.

Age of cars : 2 4 6 8	(in years)
Maintainance costs : 10 20 25 30	(Rs in 100's)

[HP Univ., MBA, 1994]

- 11.22** An analyst in a certain company was studying the relationship between travel expenses in rupees (y) for 102 sales trips and the duration in days (x) of these trips. He has found that the relationship between y and x is linear. A summary of the data is given below:

$\Sigma x = 510$; $\Sigma y = 7140$; $\Sigma x^2 = 4150$; $\Sigma xy = 54,900$, and $\Sigma y^2 = 7,40,200$

- Estimate the two regression equations from the above data.
- A given trip takes seven days. How much money should a salesman be allowed so that he will not run short of money?

- 11.23** The quantity of a raw material purchased by ABC Ltd. at specified prices during the post 12 months is given below.

Month	Price per kg (in Rs)	Quantity (in kg)	Month	Price per kg (in Rs)	Quantity (in kg)
Jan	96	250	July	112	220
Feb	110	200	Aug	112	220
March	100	250	Sept	108	200
April	90	280	Oct	116	210
May	86	300	Nov	86	300
June	92	300	Dec	92	250

- Find the regression equations based on the above data.
- Can you estimate the approximate quantity likely to be purchased if the price shoots up to Rs 124 per kg?
- Hence or otherwise obtain the coefficient of correlation between the price prevailing and the quantity demanded.

- 11.24** With ten observations on price (x) and supply (y), the following data were obtained (in appropriate units): $\Sigma x = 130$, $\Sigma y = 220$, $\Sigma x^2 = 2288$, $\Sigma y^2 = 5506$, $\Sigma xy = 3467$. Obtain the line of regression of y on x and estimate the supply when the price is 16 units. Also find out the standard error of the estimate.

- 11.25** Data on the annual sales of a company in lakhs of rupees over the past 11 years is shown below. Determine a suitable straight line regression model $y = \beta_0 + \beta_1 x + \epsilon$ for the data. Also calculate the standard error of regression of y for values of x .

Year : 1978 79 80 81 82 83 84 85 86 87 88
sales: 1 5 4 7 10 8 9 13 14 13 18

From the regression line of y on x , predict the values of annual sales for the year 1989.

- 11.26** Find the equation of the least squares line fitting the following data:

x: 1 2 3 4 5
y: 2 6 5 3 4

Calculate the standard error of estimate of y on x .

- 11.27** The following data relating to the number of weeks of experience in a job involving the wiring of an electric motor and the number of motors rejected during the past week for 12 randomly selected workers.

Workers	Experience (weeks)	No. of Rejects
1	2	26
2	9	20
3	6	28
4	14	16
5	8	23
6	12	18
7	10	24
8	4	26
9	2	38
10	11	22
11	1	32
12	8	25

- (a) Determine the linear regression equation for estimating the number of components rejected given the number of weeks of experience. Comment on the relationship between the two variables as indicated by the regression equation.
- (b) Use the regression equation to estimate the number of motors rejected for an employee with 3 weeks of experience in the job.
- (c) Determine the 95 per cent approximate prediction interval for estimating the number of motors rejected for an employee with 3 weeks of experience in the job, using only the standard error of estimate.
- 11.28** A financial analyst has gathered the following data about the relationship between income and investment in securities in respect of 8 randomly selected families:

Income : 8 12 9 24 43 37 19 16
(Rs in 1000's)

Per cent invested
in securities : 36 25 33 15 28 19 20 22

- (a) Develop an estimating equation that best describes these data.
- (b) Find the coefficient of determination and interpret it.
- (c) Calculate the standard error of estimate for this relationship.
- (d) Find an approximate 90 per cent confidence interval for the percentage of income invested in securities by a family earning Rs 25,000 annually.

[Delhi Univ., MFC, 1997]

Hints and Answers

11.15 $\bar{x} = \Sigma x / 8 = 21 / 8 = 2.625$; $\bar{y} = \Sigma y / 8 = 4 / 8 = 0.50$

$$b_{yx} = \frac{n \sum d_x d_y - (\sum d_x)(\sum d_y)}{n \sum d_x^2 - (\sum d_x)^2} = \frac{8 \times 30 - (-3)(-12)}{8 \times 45 - (-1)^2} = 0.568;$$

$$d_x = x - 3; \quad d_y = y - 3.$$

Regression equation:

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\text{or } y - 0.5 = 0.568(x - 2.625)$$

$$y = -0.991 + 0.568x$$

$$(b) b_{xy} = \frac{n \sum d_x d_y - (\sum d_x)(\sum d_y)}{n \sum d_y^2 - (\sum d_y)^2} = \frac{8 \times 30 - (-3)(-12)}{8 \times 84 - (-12)^2} = 0.386$$

Regression equation:

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$\text{or } x - 2.625 = 0.386(y - 5)$$

$$x = 0.695 + 0.386y$$

- 11.16** Let x = rainfall y = production by y . The expected yield corresponding to a rainfall of 40 inches is given by regression equation of y on x .

- 11.29** A financial analyst obtained the following information relating to return on security A and that of market M for the past 8 years:

Year :	1	2	3	4	5	6	7	8
Return A :	10	15	18	14	16	16	18	4
Market M :	12	14	13	10	9	13	14	7

- (a) Develop an estimating equation that best describes these data.
- (b) Find the coefficient of determination and interpret it.
- (c) Determine the percentage of total variation in security return being explained by the return on the market portfolio.

- 11.30** The equation of a regression line is

$$\hat{y} = 50.506 - 1.646x$$

and the data are as follows:

x:	5	7	11	12	19	25
y:	47	38	32	24	22	10

Solve for residuals and graph a residual plot. Do these data seem to violate any of the assumptions of regression?

- 11.31** Graph the following residuals and indicate which of the assumptions underlying regression appear to be in jeopardy on the basis of the graph:

x :	13	16	27	29	37	47	63
y - \hat{y} :	-11	-5	-2	-1	6	10	12

Given $\bar{y} = 50$, $\sigma_y = 10$, $\bar{x} = 30$, $\sigma_x = 5$, $r = 0.8$

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x});$$

$$y - 50 = 0.8 \frac{10}{5} (x - 30)$$

$$y = 2 + 1.6x$$

For $x = 40$, $y = 2 + 1.6(40) = 66$ quintals.

- 11.17** Let x = age of wife y = age of husband.

Given $\bar{x} = 25$, $\bar{y} = 22$, $\sigma_x = 4$, $\sigma_y = 5$, $r = 0.8$

- (a) Regression equation of x on y

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$x - 25 = 0.8 \frac{4}{5} (y - 22)$$

$$x = 10.92 + 0.64y$$

When age of wife is $y = 16$; $x = 10.92 + 0.64(16) = 22$ approx.(husband's age)

- (b) Left as an exercise

11.18 (a) Regression equation of x on y

$$x - \bar{x} = b_{xy} (y - \bar{y}) = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$x - 20 = 0.8 \frac{5}{25} (y - 120)$$

$$x = 0.8 + 0.16y$$

Regression equation of y on x

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$y - 120 = 0.8 \frac{25}{5} (x - 20)$$

$$y = 40 + 4x$$

(b) When advertisement expenditure is of Rs 25 crore, likely sales is

$$y = 40 + 4x = 40 + 4(25) = 140 \text{ crore.}$$

(c) For $y = 150$, $x = 0.8 + 0.16y = 0.8 + 0.16(150) = 24.8$

11.19 Let x = marks in Statistics and y = marks in Accountancy,

$$\text{Given: } 3y - 5x + 180 = 0$$

$$\text{or } x = (3/5)y + (180/5)$$

$$\text{For } y = 44, x = (3/5) \times 44 + (180/5) = 62.4$$

Regression coefficient of x on y , $b_{xy} = 3/5$

Coefficient of regression

$$b_{xy} = r \frac{\sigma_x}{\sigma_y} = \frac{\sqrt{9}}{\sqrt{16}} \text{ (given)}$$

$$\text{or } \frac{3}{5} = r \frac{\sqrt{9}}{\sqrt{16}} \text{ or } \frac{3}{5} = \frac{3r}{4}$$

$$\text{Hence } 3r = 2.4 \text{ or } r = 0.8$$

11.20 Let x = years of service and y = income.

(a) Regression equation of y on x

$$b_{yx} = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{n \Sigma x^2 - (\Sigma x)^2} = \frac{7 \times 469 - 56 \times 56}{7 \times 476 - (56)^2} = 0.75$$

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 8 = 0.75(x - 8)$$

$$y = 2 + 0.75x$$

(b) When $x = 13$ years, the average income would be

$$y = 2 + 0.75x = 2 + 0.75(13) = \text{Rs } 11,750$$

11.21 Let x = age of cars and y = maintainance costs.

The regression equation of y on x

$$\bar{x} = \Sigma x/n = 20/4 = 5; \quad \bar{y} = \Sigma y/n = 85/4 = 21.25$$

$$\text{and } b_{yx} = \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2} = \frac{7 \times 490 - 20 \times 85}{7 \times 120 - (20)^2} = 3.25$$

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 21.25 = 3.25(x - 5)$$

$$y = 5 + 3.25x$$

11.22 $\bar{x} = \Sigma x/n = 510/102 = 5; \quad \bar{y} = \Sigma y/n = 7140/102 = 70$

Regression coefficients:

$$b_{xy} = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n \Sigma y^2 - (\Sigma y)^2}$$

$$= \frac{102 \times 54900 - 510 \times 7140}{102 \times 740200 - (7140)^2} = 0.08$$

$$b_{yx} = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n \Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{102 \times 54900 - 510 \times 7140}{102 \times 4150 - (510)^2} = 12$$

Regression lines:

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 5 = 0.08(y - 70) \text{ or } x = 0.08y - 0.6$$

$$\text{and } y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 70 = 12(x - 5) \text{ or } y = 12x + 10$$

$$\text{When } x = 7, \quad \bar{y} = 12 \times 7 + 10 = 94$$

11.23 Let price be denoted by x and quantity by y

$$\bar{x} = \Sigma x/n = 1200/12 = 100;$$

$$\bar{y} = \Sigma y/n = 2980/12 = 248.33$$

(a) Regression coefficients:

$$b_{xy} = \frac{n \Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n \Sigma d_y^2 - (\Sigma d_y)^2} = -0.26$$

$$b_{yx} = \frac{n \Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n \Sigma d_x^2 - (\Sigma d_x)^2} = -3.244$$

Regression lines:

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 100 = -0.26(y - 248.33)$$

$$\text{or } x = -0.26y + 164.56$$

$$\text{and } y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 248.33 = -3.244(x - 100)$$

$$y = -3.244x + 572.73$$

(b) For $x = 124$,

$$y = -3.244 \times 124 + 572.73 = 170.474$$

11.24 (a) Regression line of y on x is given by

$$\text{Given } \bar{y} = \Sigma y/n = 220/10 = 22;$$

$$\bar{x} = \Sigma x/n = 130/10 = 13$$

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 22 = 1.015(x - 13)$$

$$y = 8.805 + 1.015x$$

$$b_{yx} = \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2} = \frac{10 \times 3467 - 130 \times 220}{10 \times 2288 - (130)^2}$$

$$= \frac{34670 - 28600}{22880 - 16900} = \frac{6070}{5980} = 1.015$$

(b) When $x = 16$,

$$y = 8.805 + 1.015(16) = 25.045$$

$$(c) S_{yx} = S_y \sqrt{1 - r^2} = 8.16 \sqrt{1 - (0.9618)^2} = 2.004$$

11.25 Take years as $x = -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5$, where 1983 = 0. The regression equation is

$$\hat{y} = 9.27 + 1.44x$$

$$\text{For } x = 1989, \quad \hat{y} = 9.27 + 1.44(6) = 17.91$$

$$S_{yx} = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}} = \sqrt{\frac{21.2379}{10}} = 1.4573$$

11.26 $\bar{x} = \Sigma x/n = 15/5 = 3$, $\bar{y} = \Sigma y/n = 20/5 = 4$

The regression equation is:

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 4 = 0.7(x - 3) \text{ or } \hat{y} = 1.9 + 0.7x$$

Standard error of estimate,

$$S_{yx} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n-2}} = \sqrt{\frac{5.1}{3}} = 1.303$$

11.27 (a) $b = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n(\bar{x})^2} = \frac{2048 - 12(7.67)(24.83)}{876 - 12(7.67)^2} = -1.40$

$$a = \bar{y} - b\bar{x} = 24.83 - (-1.40)(7.67) = 35.57$$

Thus $\hat{y} = a + bx = 35.57 - 1.40x$

Since $b = -1.40$, it indicate an inverse (negative)

relationship between weeks of experience (x) and the number of rejects (y) in the sample week

(b) For $x = 3$, we have $\hat{y} = 35.57 - 1.40(3) \approx 31$

$$(c) S_{yx} = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n-2}}$$

$$= \sqrt{\frac{7,798 - (35.57)(298) - 1.40(2048)}{12-2}}$$

$$= 2.56$$

95 per cent approximate prediction interval

$$\hat{y} \pm t_{df} S_{yx} = 31.37 \pm 2.228 (2.56)$$

= 25.67 to 37.07 or 26 to 37 rejects.

11.28 4.724; -0.983; -0.399, -6.753, 2.768, 0.644

11.29 Error term non-independent.

Case Studies

Case 11.1: Made in China

The phrase 'made in China' has become an issue of concern in the last few years, as Indian Companies try to protect their products from overseas competition. In these years a major trade imbalance in India has been caused by a flood of imported goods that enter the country and are sold at lower price than comparable Indian made goods. One prime concern is the electronic goods in which total imported items have steadily increased during the year 1990s to 2004s. The Indian companies have been worried on complaints about product quality, worker layoffs, and high prices and has spent millions in advertising to produce electronic goods that will satisfy consumer demands. Have these companies been successful in stopping the flood these imported goods purchased by Indian consumers? The given data represent the volume of imported goods sold in India for the years 1999-2004. To simplify the analysis, we have coded the year using the coded variable $x = \text{Year 1989}$.

Year	$x = \text{Year 1989}$	Volume of Import (in Rs billion)
1989	0	1.1
1990	1	1.3
1991	2	1.6
1992	3	1.6
1993	4	1.8
1994	5	1.4
1995	6	1.6
1996	7	1.5
1997	8	2.1
1998	9	2.0
1999	10	2.3
2000	11	2.4
2001	12	2.3
2002	13	2.2
2003	14	2.4
2004	15	2.4

Questions for Discussion

- Find the least-squares line for predicting the volume of import as a function of year for the years 1989-2000.
- Is there a significant linear relationship between the volume of import and the year?
- Predict the volume of import of goods using 95% prediction intervals for each of the years 2002, 2003 and 2004.
- Do the predictions obtained in Step 4 provide accurate estimates of the actual values observed in these years? Explain.
- Add the data for 1989-2004 to your database, and recalculate the regression line. What effect have the new data points had on the slope? What is the effect of SSE?
- Given the form of the scattered diagram for the years 1989-2004, does it appear that a straight line provides an accurate model for the data? What other type of model might be more appropriate?

This page is intentionally left blank.

I have but one lamp by which my feet are guided, and that is the lamp of experience. I know of no way of judging the future but the past.

—Patrick Henry

The penguin flies backwards because he does not care to see where he's going, but wants to see where he's been.

—Fred Allen

Forecasting and Time Series Analysis

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- understand the pattern of the historical data and then extrapolate the pattern into the future.
- understand the different approaches to forecasting that can be applied in business.
- gain a general understanding of time-series forecasting techniques.
- learn how to decompose time-series data into their various components and to forecast by using decomposition techniques.

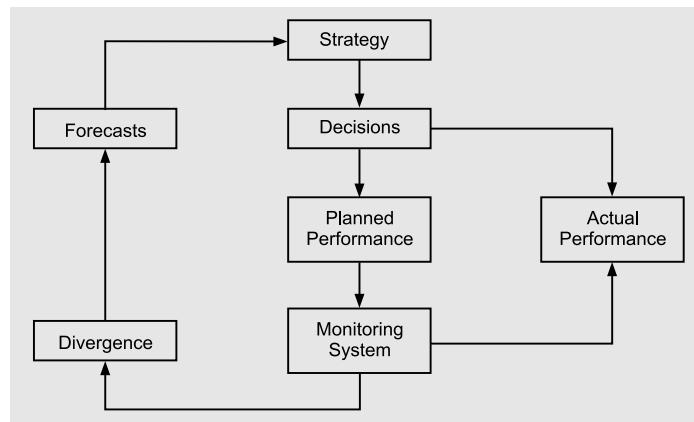
12.1 INTRODUCTION

The increasing complexity of the business environment together with changing demands and expectations, implies that every organization needs to know the future values of their key decision variables. Forecasting takes the historical data and project them into the future to predict the occurrence of uncertain events. This may help organizations to assess the future consequences of existing decisions and to evaluate the consequences of decisions (actions or strategies). For example, inventory is ordered without certainty of future sales; new equipment is purchased despite uncertainty about the demand for products; investments are made without knowing profits in future; alternative staff mix is made without knowing the increase in the level of service that can be provided, and so on.

Forecasting is essential to make reliable and accurate estimates of what will happen in the future in the face of uncertainty. A flow chart of forecasts and the decision-making process is shown in Fig. 12.1. In general, the decisions are influenced by the chosen strategy with regard to an organization's future priorities and activities. Once decisions are taken, the consequences are measured in terms of expectation to achieve the desired products/services levels.

Decisions also get influenced by the additional information obtained from the forecasting method used. Such information and the perceived accuracy of the forecasts may also affect the strategy formulation of an organization. Thus an organization needs

Figure 12.1
Decision-Making Process and Forecasts



to establish a monitoring system to compare planned performance with the actual. Divergence, if any, and no matter what the cause of such divergence between the planned and actual performance, should be fed back into the forecasting process, to generate new forecasts. A few objectives of forecasting are as follows:

- (i) The creation of plans of action, because it is not possible to evolve a system of business control without an acceptable system of forecasting.
- (ii) Monitoring of the continuing progress of action plans based on forecasts.
- (iii) The forecast provides a warning system of the critical factors to be monitored regularly because they might drastically affect the performance of the plan.

12.2 TYPES OF FORECASTS

The objectives of any organization are facilitated by a number of different types of forecasts. These may be related to cash flows, operating budgets, personnel requirement, inventory levels, and so on. However, a broad classification of the types of forecasts is as follows:

Demand Forecasts These are concerned with the predictions of demand for products or services. These forecasts facilitate in formulating material and capacity plans and serve as inputs to financial, marketing, and personnel planning. The forecast itself may be generated in a number of ways, many of which depend heavily upon sales and marketing information.

Environmental Forecasts These are concerned with the social, political, and economic environment of the state and/or the country. Environmental concerns, such as pollution control, are much better managed from an anticipatory rather than an after-the-fact standpoint. Economic forecasts are valuable because they help in predicting inflation rates, money supplies, operating budget, and so on.

Technological Forecasts These are concerned with new developments in existing technologies as well as the development of new technologies. They have become increasingly important to major firms in the computer, aerospace, nuclear, and many other technologically advanced industries.

12.3 TIMING OF FORECASTS

Forecasts are usually classified according to time period and use. The three categories of forecasts are:

Short-Range Forecast This has a time span of upto one year but is typically less than three months. It is normally used in planning purchasing for job scheduling, work force levels, job assignments, production levels, and the like.

Medium-Range Forecast This has a time span from one to three years (typically 3 months to one year). It is used for sales planning, production planning, cash budgeting, and so on.

Long-Range Forecast This has a time span of three or more years. It is used for designing and installing new plants, facility location, capital expenditures, research and development, and so on.

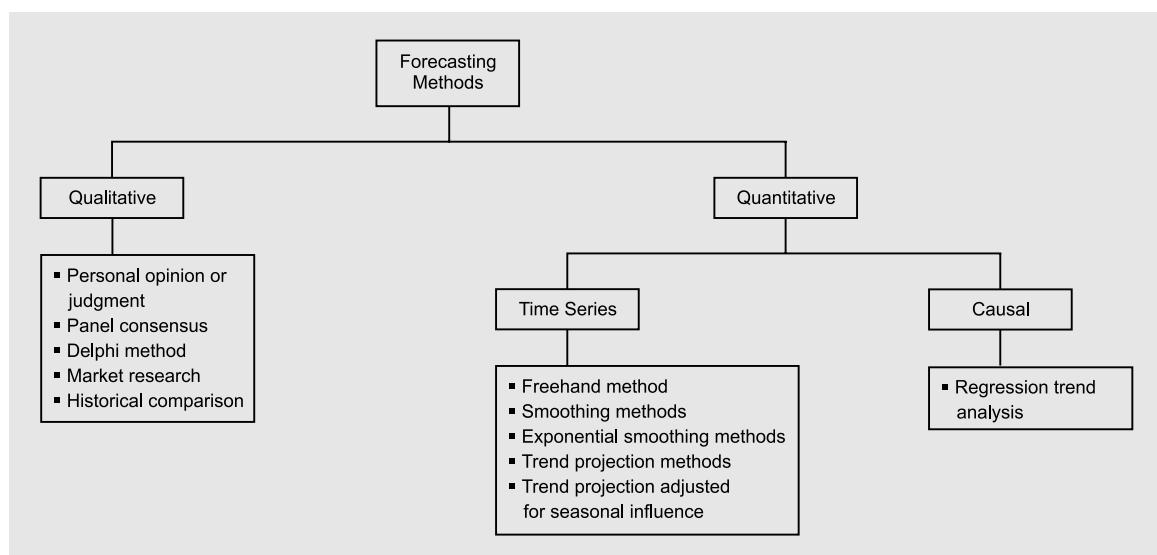
The medium and long-range forecasts differ from short-range forecast on account of following three features:

- (i) Medium and long-range forecasts deal with more comprehensive issues and support management decisions regarding design and development of new products, plants, and processes.
- (ii) Mathematical techniques such as moving averages, exponential smoothing, and trend extrapolation are used for short-range forecasts.
- (iii) The short-range forecasts tend to be more accurate than long-range forecasts. For example, sales forecasts need to be updated regularly in order to maintain their value. After each sales period, the forecast should be reviewed and revised.

12.4 FORECASTING METHODS

Forecasting methods may be classified as either quantitative or qualitative (opinion or judgmental). Figure 12.2 provides an overview of the types of forecasting methods.

Figure 12.2
Forecasting Methods



12.4.1 Quantitative Forecasting Methods

These methods can be used when

- (i) past information about the variable being forecast is available,
- (ii) information can be quantified, and
- (iii) a reasonable assumption is that the pattern of the past will continue into the future.

The quantitative methods of forecasting are further classified into two categories:

Time Series Forecasting Methods A time series is a set of measurements of a variable that are ordered through time. The time variable does not fluctuate arbitrarily. It moves uniformly always in the same direction, from past to future. Thus we can exercise some freedom of choice as to the times at which observations can be made. The time-series data are gathered on a given variable characteristic over a period of time at regular intervals.

The time series forecasting methods attempt to account for changes over a period of time at regular intervals by examining patterns, cycles or trends to predict the outcome for a future time period.

Causal forecasting methods: Forecasting methods that relate a time-series to other variables which are used to explain cause and effect relationship.

Causal Forecasting Methods These methods are based on the assumptions that the variable value which we intend to forecast has a cause-effect relationship with one or more other variables. A linear regression analysis which depends upon the causal relationship or interaction of two or more variables is called causal forecasting method.

12.4.2 Qualitative Forecasting Methods

These methods consist of collecting the opinions and judgments of individuals who are expected to have the best knowledge of current activities or future plans of the organization. For example, knowledge of demand trend and customer plans are often known to marketing executives or product managers. Through regular contact with customers, the marketing and sales personnel are presumably familiar with individual customers or retail market segment. Management usually maintains broader market information on trends by product line, geographic area, customer groups, and so on.

Qualitative forecasting methods have the advantage that they can incorporate subjective experience as inputs along with objective data. It is the human brain that permits assimilation of all types of information and the ultimate issuance of a prediction.

Since each human being has different knowledge, experience, and perspective of reality, intuitive forecasts are likely to differ from one individual to another. Furthermore, the less they are based upon fact and quantified data, the less they lend themselves to analysis and resolution of differences of opinion. The quantification of data gives them a more precise meaning than words which are inexact and are capable of being misunderstood. Also, if the forecasts prove to be inaccurate there is an objective basis for improvement the next time around.

A number of approaches fall under qualitative methods, and these are as follows:

Personal Opinion In this approach of forecasting, an individual does some forecast of the future based on his or her own judgment or opinion without using a formal quantitative model. Such an assessment can be relatively reliable and accurate. This approach is usually recommended when conditions in the past are not likely to hold in the future. For instance, getting an assessment of whether inventory levels are likely to last until the next replenishment; whether a machine will require repair in the next month, and so on.

Panel Consensus To reduce the prejudices and ignorance that may arise in the individual judgment, it is possible to develop consensus among group of individuals. Such a panel of individuals is encouraged to share information, opinions, and assumptions (if any) to predict future value of some variable under study.

The disadvantage of this method is that it is dependent on group dynamics and frequently requires a facilitator or convenor to coordinate the process of developing a consensus.

Delphi method: A quantitative forecasting method that obtains forecasts through group consensus.

Delphi Method This method is very similar to the panel consensus approach. It uses the collective experience and judgment of a group of experts. In this method, experts may be located in different places and never meet and typically do not know other group members. Each expert is given a questionnaire to complete relating to the area under investigation. A summary is then prepared from all the questionnaires and a copy of it is sent to each expert for revision of responses to the question included in the questionnaire in the light of the summary results. This process of updating the summary results is repeated until the desirable consensus is reached. This method produces a narrow range of forecasts rather than a single view of the future.

Market Research This method is used to collect data based on well-defined objectives and assumptions about the future value of a variable. In this method, a questionnaire is prepared to distribute among respondents. A summary of responses to questions in the questionnaire is prepared to develop survey results.

Historical Comparison Once the data are arranged chronologically, the time-series approach facilitates comparison between one time period and the next. It provides a scientific basis for making comparisons by studying and isolating the effects of various influencing factors on the patterns of variable values. It also helps in making regional comparison amongst data collected on the basis of time.

12.5 STEPS OF FORECASTING

Regardless of the method used to forecast, the following steps are followed:

1. Define objectives and the policies to be achieved, that is, what we are trying to obtain by the use of the forecast. The purpose of forecasting is to make use of the best available present information to guide future activities towards organization's objectives.
2. Select the variables of interest such as capital investment, employment level, inventory level, purchasing of new equipment, which are to be forecasted.
3. Determine the time horizon—short, medium, or long term—of the forecast in order to predict changes which will probably follow the present level of activities.
4. Select an appropriate forecasting model to make projections of the future in accordance to the reasons of past changes which have taken place.
5. Collect the relevant data needed to make the forecast.
6. Make the forecast and implement the results.

These steps present a systematic way of initiating, designing, and implementing a forecasting system. If a particular system is used regularly to generate forecasts, then data should be collected in a routine manner so that computations used to make the forecast can be done automatically using a computer.

12.6 TIME SERIES ANALYSIS

A time series is a set of numerical values of some variable obtained at regular period over time. The series is usually tabulated or graphed in a manner that readily conveys the behaviour of the variable under study. Figure 12.3 presents the export of cement (in tonnes) by a cement company between 1994 and 2004. The graph suggests that the series is time dependent. The management of the company is interested in determining how the series is dependent on time and in developing a means of predicting future levels with some degree of reliability. The nature of the time dependence is often analysed by decomposing the time series into its components.

Year	Export (tonnes)
1994	2
1995	3
1996	6
1997	10
1998	8
1999	7
2000	12
2001	14
2002	14
2003	18
2004	19

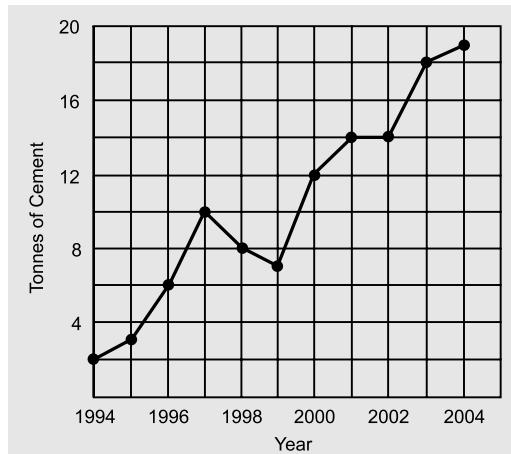


Figure 12.3
Export of Cement

12.6.1 Objectives of Time Series Analysis

1. The assumption underlying time series analysis is that the future will look like the past, that is, the various factors which have already influenced the patterns of change in the value of the variable under study will continue to do so in more or less the same manner in the future. In other words, some underlying pattern exists in historical data. Thus one of the objective of time-series analysis is to identify the pattern and isolate the influencing factors (or effects) for prediction purposes as well as for future planning and control.

2. The review and evaluation of progress made on the basis of a plan are done on the basis of time-series data. For example the progress of our Five-Year Plans is judged by the annual growth rates in the Gross National Product (GNP). Similarly the evaluation of our policy of controlling inflation and price rise is done by the study of various price indices which are based on the analysis of time-series.

12.6.2 Time Series Patterns

We assume that time series data consist of an underlying pattern accompanied by random fluctuations. This may be expressed in the following form:

Actual value of the variable at time t = Mean value of the variable at time t + Random deviation from mean value

$$\text{variable at time } t = \text{Pattern} + e$$

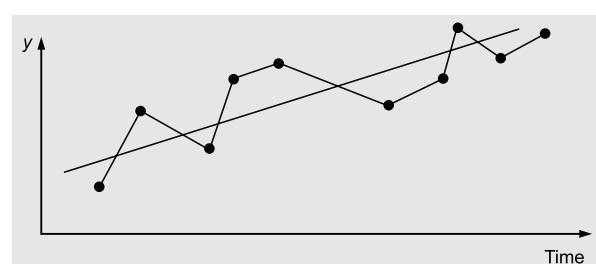
where \hat{y} is the forecast variable at period t ; pattern is the mean value of the forecast variable at period t and represents the underlying pattern, and e is the random fluctuation from the pattern that occurs of the forecast variable at period t .

12.6.3 Components of a Time Series

The **time-series** data contain four components: *trend*, *cyclical*, *seasonality* and *irregularity*. Not all time-series have all these components. Figure 12.4 shows the effects of these time-series components over a period of time.

Trend Sometimes a time-series displays a steady tendency of either upward or downward movement in the average (or mean) value of the forecast variable y over time. Such a tendency is called a trend. When observations are plotted against time, a straight line describes the increase or decrease in the time series over a period of time.

Cycles An upward and downward movement in the variable value about the trend time over a time period are called cycles. A business cycle may vary in length, usually more than a year but less than 5 to 7 years. The movement is through four phases: from *peak* (prosperity) to *contradiction* (recession) to *trough* (depression) to *expansion* (recovery or growth) as shown in Fig. 12.4 .



Time-series: A set of observations measured at successive points in time or over successive periods of time.

Trend: A type of variation in time-series that reflects a long-term movement in time-series over a long period of time.

Cyclical variation: A type of variation in time-series, in which the value of the variable fluctuates above and below a trend line and lasting more than one year.

Figure 12.4
Time-series Effects

Seasonal variation: A type of variation in time-series that shows a periodic pattern of change in time-series within a year; patterns tend to be repeated from year to year.

Irregular variation: A type of variation in time-series that reflects the random variation of the time-series values which is completely unpredictable.

Seasonal It is a special case of a cycle component of time series in which fluctuations are repeated usually within a year (e.g. daily, weekly, monthly, quarterly) with a high degree of regularity. For example, average sales for a retail store may increase greatly during festival seasons.

Irregular Irregular variations are rapid changes or *bleeps* in the data caused by short-term unanticipated and non-recurring factors. Irregular fluctuations can happen as often as day to day.

12.7 TIME SERIES DECOMPOSITION MODELS

The analysis of time series consists of two major steps:

1. Identifying the various factors or influences which produce the variations in the time series, and
2. Isolating, analysing and measuring the effect of these factors independently, by holding other things constant.

The purpose of decomposition models is to break a time series into its components: Trend (T), Cyclical (C), Seasonality (S), and Irregularity (I). Decomposition of time series aims to isolate influence of each of the four components on the actual series so as to provide a basis for forecasting. There are many models by which a time series can be analysed; two models commonly used for decomposition of a time series are discussed below.

12.7.1 Multiplicative Model

The actual values of a time series, represented by Y can be found by multiplying four components at a particular time period. The effect of four components on the time series is interdependent. The multiplicative time series model is defined as:

$$Y = T \times C \times S \times I \leftarrow \text{Multiplicative model}$$

The multiplicative model is appropriate in situations where the effect of C , S , and I is measured in relative sense and is not in absolute sense. The geometric mean of C , S , and I is assumed to be less than one. For example, let the actual sales for period of 20 months be $Y_{20} = 423.36$. Further let this value be broken down into its components as: trend component (mean sales) 400; effect of current cycle (0.90) which decreases sales by 10 per cent; seasonality of the series (1.20) that increases sales by 20 per cent. Thus besides the random fluctuation, the expected value of sales for this period is: $400 \times 0.90 \times 1.20 = 432$. If the random factor decreases sales by 2 per cent in this period, then the actual sales value will be $432 \times 0.98 = 423.36$.

12.7.2 Additive Model

In this model, it is assumed that the effect of various components can be estimated by adding the various components of a time-series. It is stated as:

$$Y = T + C + S + I \leftarrow \text{Additive model}$$

Here C , S , and I are absolute quantities and can have positive or negative values. It is assumed that these four components are independent of each other. However, in real-life time series data this assumption does not hold good.

Conceptual Questions 12A

1. Briefly describe the steps that are used to develop a forecasting system.
2. What is forecasting? Discuss in brief the various theories and methods of business forecasting.
[Delhi Univ., MBA, 2001]
3. For what purpose do we apply time series analysis to data collected over a period of time?
4. How can one benefit from determining past patterns?
5. What is the difference between a causal model and a time series model?
6. What is a judgmental forecasting model, and when is it appropriate?
7. Explain clearly the different components into which a time series may be analysed. Explain any method for isolating trend values in a time series.
8. Explain what you understand by time series. Why is time-series considered to be an effective tool of forecasting?
9. Explain briefly the additive and multiplicative models of time series. Which of these models is more popular in practice and why? [Osmania Univ., MBA, 1998]
10. Identify the four principal components of a time-series and explain the kind of change, over time, to which each applies.
11. What is the advantage of reducing a time series into its four components?
12. Despite great limitations of statistical forecasting, forecasting techniques are invaluable to the economist, the businessman, and the government. Explain.
13. (a) Why are forecasts important to organizations?
(b) Explain the difference between the terms: seasonal variation and cyclical variation.
(c) Give reasons why the seasonal component in the time-series is not constant? Give examples where you believe the seasonality may change.
14. Identify the classical components of a time series and indicate how each is accounted for in forecasting.

12.8 QUANTITATIVE FORECASTING METHODS

The quantitative forecasting methods fall into two general categories:

- Time series methods
- Causal methods

The **time series methods** are concerned with taking some observed historical pattern for some variable and projecting this pattern into the future using a mathematical formula. These methods do not attempt to suggest why the variable under study will take some future value. This limitation of the time-series approach is taken care by the application of a causal method. The **causal method** tries to identify factors which influence the variable in some way or cause it to vary in some predictable manner. The two causal methods, regression analysis and correlation analysis, have already been discussed previously.

A few time series methods such as *freehand curves* and *moving averages* simply describe the given data values, while other methods such as *semi-average* and *least squares* help to identify a trend equation to describe the given data values.

12.8.1 Freehand Method

A freehand curve drawn smoothly through the data values is often an easy and, perhaps, adequate representation of the data. From Fig. 12.3, it appears that a straight line connecting the 1994 and 2004 exports volumes is a fairly good representation of the given data.

The forecast can be obtained simply by extending the trend line. A trend line fitted by the freehand method should confirm to the following conditions:

- (i) The trend line should be smooth—a straight line or mix of long gradual curves.
- (ii) The sum of the vertical deviations of the observations above the trend line should equal the sum of the vertical deviations of the observations below the trend line.
- (iii) The sum of squares of the vertical deviations of the observations from the trend line should be as small as possible.
- (iv) The trend line should bisect the cycles so that area above the trend line should be equal to the area below the trend line, not only for the entire series but as much as possible for each full cycle.

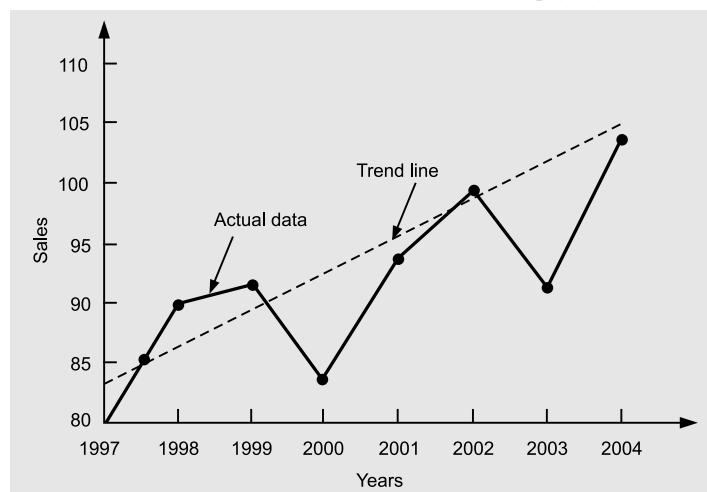
Example 12.1: Fit a trend line to the following data by using the freehand method.

Year	: 1997	1998	1999	2000	2001	2002	2003	2004
Sales turnover :	80	90	92	83	94	99	92	104.

(Rs in lakh)

Solution: Figure 12.5 presents the freehand graph of sales turnover (Rs in lakh) from 1997 to 2004. Forecast can be obtained simply by extending the trend line

Figure 12.5
Graph of Sales Turnover



Limitations of freehand method

- (i) This method is highly subjective because the trend line depends on personal judgment and therefore what happens to be a good-fit for one individual may not be so for another.
- (ii) The trend line drawn cannot have much value if it is used as a basis for predictions.
- (iii) It is very time-consuming to construct a freehand trend if a careful and conscientious job is to be done.

12.8.2 Smoothing Methods

The objective of smoothing methods is to smoothen out the random variations due to irregular components of the time series and thereby provide us with an overall impression of the pattern of movement in the data over time. In this section, we shall discuss three smoothing methods:

- (i) Moving averages
- (ii) Weighted moving averages
- (iii) Semi-averages

The data requirements for the techniques to be discussed in this section are minimal and these techniques are easy to use and understand.

Moving Averages

If we attempt to observe the movement of some variable values over a period of time and try to project this movement into the future, then it is essential to smooth out first the irregular pattern in the historical values of the variable, and later use this as the basis for a future projection. This can be done by using the technique of **moving averages**.

This method is a subjective method and depends on the length of the period chosen for calculating moving averages. To remove the effect of cyclical variations, the period chosen should be an integer value that corresponds to or is a multiple of the estimated average length of a cycle in the series.

The moving averages which serve as an estimate of the next period's value of a variable given a period of length n is expressed as:

$$\text{Moving average, } MA_{t+1} = \frac{\Sigma\{D_t + D_{t-1} + D_{t-2} + \dots + D_{t-n+1}\}}{n}$$

where t = current time period

D = actual data which is exchanged each period

n = length of time period

In this method, the term 'moving' is used because it is obtained by summing and averaging the values from a given number of periods, each time deleting the oldest value and adding a new value.

The major *advantage* of a moving average is the opportunity it provides to focus on the long-term trend (and cyclical) movements in a time series without the obscuring effect of short-term 'noise' influences.

The *limitation* of this method is that it is highly subjective and dependent on the length of period chosen for constructing the averages. Moving averages have the following three limitations:

- (i) As the size of n (the number of periods averaged) increases, it smoothens the variations better, but it also makes the method less sensitive to real changes in the data.
- (ii) It is difficult to choose the optimal length of time for which to compute the moving average. Moving averages can not be found for the first and last $k/2$ periods in a k -period moving average.
- (iii) Moving averages cannot pick-up trends very well. Since these are averages, it will always stay within past levels and will not predict a change to either a higher or lower level.

Moving averages: A quantitative method of forecasting or smoothing a time-series by averaging each successive group of data values.

- (iv) It causes a loss of information (data values) at either end of the original time series.
- (v) Moving averages do not usually adjust for such time-series effects as trend, cycle or seasonality.

Example 12.2: Shown is production volume (in '000 tonnes) for a product. Use these data to compute a 3-year moving average for all available years. Also determine the trend and short-term error.

Year	Production (in '000 tonnes)	Year	Production (in '000 tonnes)
1995	21	2000	22
1996	22	2001	25
1997	23	2002	26
1998	25	2003	27
1999	24	2004	26

Solution: The first average is computed for the first 3 years as follows:

$$\text{Moving average (year 1-3)} = \frac{21 + 22 + 23}{3} = 22$$

The first 3-year moving average can be used to forecast the production volume in fourth year, 1998. Because 25,000 tonnes production was made in 1998, the error of the forecast is $\text{Error}_{1998} = 25,000 - 22,000 = 3000$ tonnes.

Similarly, the moving average calculation for the next 3 years is:

$$\text{Moving average (year 2-4)} = \frac{22 + 23 + 25}{3} = 23.33$$

A complete summary of 3-year moving average calculations is given in Table 12.1.

Table 12.1 Calculation of Trend and Short-term Fluctuations

Year	Production y	3-Year Moving Total	3-Yearly Moving Average (Trend values)	Forecast Error $(y - \hat{y})$
1995	21	—	—	—
1996	22	$\rightarrow (21 + 22 + 23) = 66$	$66/3 = 22.00$	0
1997	23	$\rightarrow (22 + 23 + 25) = 70$	$70/3 = 23.33$	-0.33
1998	25	$\rightarrow (23 + 25 + 24) = 72$	$72/3 = 24.00$	1.00
1999	24	71	23.67	0.33
2000	22	71	23.67	-1.67
2001	25	73	24.33	0.67
2002	26	78	26.00	0
2003	27	$\rightarrow (26 + 27 + 26) = 79$	$79/3 = 26.33$	0.67
2004	26	—	—	—

Odd and Even Number of Years When the chosen period of length n is an odd number, the moving average period is centred on i (middle period in the consecutive sequence of n periods). For instance with $n = 5$, $\text{MA}_3(5)$ is centred on the third year, $\text{MA}_4(5)$ is centred on the fourth year..., and $\text{MA}_9(5)$ is centred on the ninth year.

No moving average can be obtained for the first $(n - 1)/2$ years or the last $(n - 1)/2$ year of the series. Thus for a 5-year moving average, we cannot make computations for the just two years or the last two years of the series.

When the chosen period of length n is an even numbers, equal parts can easily be formed and an average of each part is obtained. For example, if $n = 4$, then the first moving average M_3 (placed at period 3) is an average of the first four data values, and the second moving average M_4 (placed at period 4) is the average of data values 2 through 5. The average of M_3 and M_4 is placed at period 3 because it is an average of data values for period 1 through 5.

Example 12.3: Assume a four-year cycle and calculate the trend by the method of moving average from the following data relating to the production of tea in India:

Year	Production (million lbs)	Year	Production (million lbs)
1987	464	1992	540
1988	515	1993	557
1989	518	1994	571
1990	467	1995	586
1991	502	1996	612

[Madras, Univ., MCom, 1997]

Solution: The first 4-year moving average is:

$$MA_3(4) = \frac{464 + 515 + 518 + 467}{4} = \frac{1964}{4} = 491.00$$

This moving average is centred on the middle value, that is, the third year of the series.

Similarly,

$$MA_4(4) = \frac{515 + 518 + 467 + 502}{4} = \frac{2002}{4} = 500.50$$

This moving average is centred on the fourth year of the series.

Table 12.2 presents the data along with the computations of 4-year moving averages.

Table 12.2 Calculation of Trend and Short-term Fluctuations

Year	Production (mn lbs)	4-Yearly Moving Totals	4-Yearly Moving Average	4-Yearly Moving Average Centred
1987	464	—	—	—
1988	515	—	—	—
1989	518	1964	491.00	495.75
1990	467	2002	500.50	503.62
1991	502	2027	506.75	511.62
1992	540	2066	516.50	529.50
1993	557	2170	542.50	553.00
1994	571	2254	563.50	572.50
1995	586	2326	581.50	—
1996	612	—	—	—

Weighted Moving Averages

In moving averages, each observation is given equal importance (weight). However, it may be desired to place more weight (importance) on certain periods of time than on others. So a moving average in which some time periods are weighted differently than others is called a

Weighted moving average: A quantitative method of forecasting or smoothing a time-series by computing a weighted average of past data values; sum of weights must equal one.

weighted moving average. In such a case different values may be assigned to compute a weighted average of the most recent n values. Choice of weights is somewhat arbitrary because there is no set formula to determine them. In most cases, the most recent observation receives the most weightage, and the weight decreases for older data values.

A weighted moving average is computed as:

$$\text{Weighted moving average} = \frac{\sum(\text{Weight for period } n)(\text{Data value in period } n)}{\sum \text{Weights}}$$

Example 12.4: Vacuum cleaner sales for 12 months is given below. The owner of the supermarket decides to forecast sales by weighting the past three months as follows:

	<i>Weight Applied</i>			<i>Month</i>								
	3			Last month								
	2			Two months ago								
	1			Three months ago								
	<hr/>			<hr/>								
Months :	1	2	3	4	5	6	7	8	9	10	11	12
Actual sales : (in units)	10	12	13	16	19	23	26	30	28	18	16	14

Solution: The results of 3-month weighted average are shown in Table 12.3

$$\begin{aligned}\bar{x}_{\text{weighted}} &= 3M_{t-1} + 2M_{t-2} + 1M_{t-3} \\ &= \frac{1}{6}[3 \times \text{Sales last month} + 2 \times \text{Sales two months ago} + 1 \times \text{Sales three months ago}]\end{aligned}$$

Table 12.3 Weighted Moving Average

<i>Month</i>	<i>Actual Sales</i>	<i>Three-month Weighted Moving Average</i>
1	10	—
2	12	—
3	13	—
4	16	$\frac{1}{6}[(3 \times 13) + (2 \times 12) + (1 \times 10)] = \frac{121}{6}$
5	19	$\frac{1}{6}[(3 \times 16) + (2 \times 13) + (1 \times 12)] = \frac{141}{3}$
6	23	$\frac{1}{6}[(3 \times 19) + (2 \times 16) + (1 \times 13)] = 17$
7	26	$\frac{1}{6}[(3 \times 23) + (2 \times 19) + (1 \times 16)] = \frac{201}{2}$
8	30	$\frac{1}{6}[(3 \times 26) + (2 \times 23) + (1 \times 19)] = \frac{235}{6}$
9	28	$\frac{1}{6}[(3 \times 30) + (2 \times 26) + (1 \times 23)] = \frac{271}{2}$
10	18	$\frac{1}{6}[(3 \times 28) + (2 \times 30) + (1 \times 26)] = \frac{289}{3}$
11	16	$\frac{1}{6}[(3 \times 18) + (2 \times 28) + (1 \times 30)] = \frac{231}{3}$
12	14	$\frac{1}{6}[(3 \times 16) + (2 \times 18) + (1 \times 28)] = \frac{182}{3}$

Example 12.5: A food processor uses a moving average to forecast next month's demand. Past actual demand (in units) is shown below:

Month :	43	44	45	46	47	48	49	50	51
Actual demand :	105	106	110	110	114	121	130	128	137

- (a) Compute a simple five-month moving average to forecast demand for month 52.
- (b) Compute a weighted three-month moving average where the weights are highest for the latest months and descend in order of 3, 2, 1.

Solution: Calculations for five-month moving average are shown in Table 12.4.

Table 12.4 Five-month Moving Average

Month	Actual Demand	5-month Moving Total	5-month Moving Average
43	105	—	—
44	106	—	—
45	110	545	109.50
46	110	561	112.2
47	114	585	117.0
48	121	603	120.6
49	130	630	126.0
50	128	—	—
51	137	—	—

- (a) Five-month average demand for month 52 is

$$\frac{\sum x}{\text{Number of periods}} = \frac{114 + 121 + 130 + 128 + 137}{5} = 126 \text{ units}$$

- (b) Weighted three-month average as per weights is as follows:

$$\bar{x}_{\text{weighted}} = \frac{\sum \text{Weight} \times \text{Data value}}{\sum \text{weight}}$$

where	Month	Weight	\times	Value	=	Total
	51	3	\times	137	=	411
	50	2	\times	128	=	256
	49	1	\times	130	=	130
		6				797

$$\bar{x}_{\text{weighted}} = \frac{797}{6} = 133 \text{ units.}$$

Semi-Average Method

The semi-average method permits us to estimate the slope and intercept of the trend line quite easily if a linear function will adequately describe the data. The procedure is simply to divide the data into two parts and compute their respective arithmetic means. These two points are plotted corresponding to their midpoint of the class interval covered by the respective part and then these points are joined by a straight line, which is the required trend line. The arithmetic mean of the first part is the intercept value, and the slope is determined by the ratio of the difference in the arithmetic mean of the number of years between them, that is, the change per unit time. The resultant is a time series of the form : $\hat{y} = a + bx$. The \hat{y} is the calculated trend value and a and b are the intercept and slope values respectively. The equation should always be stated completely with reference to the year where $x = 0$ and a description of the units of x and y .

The semi-average method of developing a trend equation is relatively easy to

commute and may be satisfactory if the trend is linear. If the data deviate much from linearity, the forecast will be biased and less reliable.

Example 12.6: Fit a trend line to the following data by the method of semi-average and forecast the sales for the year 2002.

Year	Sales of Firm (thousand units)	Year	Sales of Firm (thousand units)
1993	102	1997	108
1994	105	1998	116
1995	114	1999	112
1996	110		

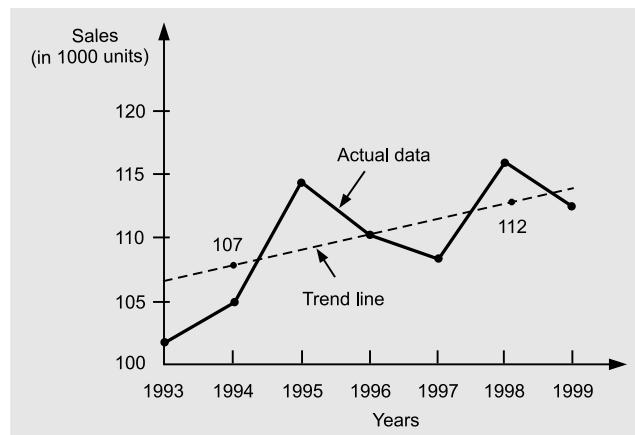
Solution: Since number of years are odd in number, therefore divide the data into equal parts (A and B) of 3 years ignoring the middle year (1996). The average of part A and B is

$$\bar{y}_A = \frac{102 + 105 + 114}{3} = \frac{321}{3} = 107 \text{ units}$$

$$\bar{y}_B = \frac{108 + 116 + 112}{3} = \frac{336}{3} = 112 \text{ units}$$

Part A is centred upon 1994 and part B on 1998. Plot points 107 and 112 against their middle years, 1994 and 1998. By joining these points, we obtain the required trend line as shown Fig. 12.6. The line can be extended and be used for prediction

Figure 12.6
Trend Line by the Method of
Semi-Average



To calculate the time-series $\hat{y} = a + bx$, we need

$$\begin{aligned} \text{Slope } b &= \frac{\Delta y}{\Delta x} = \frac{\text{change in sales}}{\text{change in year}} \\ &= \frac{112 - 107}{1998 - 1994} = \frac{5}{4} = 1.25 \end{aligned}$$

$$\text{Intercept } a = 107 \text{ units at 1994}$$

Thus, the trend line is: $\hat{y} = 107 + 1.25x$

Since 2002 is 8 year distant from the origin (1994), therefore we have

$$\hat{y} = 107 + 1.25(8) = 117$$

12.8.3 Exponential Smoothing Method

Exponential smoothing is a type of moving-average forecasting technique which weighs past data from previous time periods with exponentially decreasing importance in the forecast so that the most recent data carries more weight in the moving average. Simple exponential smoothing makes no explicit adjustment for trend effects whereas adjusted exponential smoothing does take trend effects into account (see next section for details).

Exponential smoothing method: A quantitative forecasting method that uses a weighted average of past time-series values to arrive at new time-series values.

Simple Exponential Smoothing

With simple exponential smoothing, the **forecast** is made up of the actual value for the present time period X_t multiplied by a value between 0 and 1 (the exponential smoothing constant) referred to as α (not the same as used for a Type I error) plus the product of the present time period forecast F_t and $(1 - \alpha)$. The formula is stated algebraically as follows:

$$F_{t+1} = \alpha X_t + (1 - \alpha) F_t = F_t + \alpha (X_t - F_t) \quad (12-1a)$$

where F_{t+1} = Forecast for the next time period ($t + 1$)

F_t = forecast for the present time period (t)

α = a weight called exponentially smoothing constant ($0 \leq \alpha \leq 1$)

X_t = actual value for the present time period (t)

If exponential smoothing has been used over a period of time, the forecast for F_t will have been obtained by

$$F_t = \alpha X_{t-1} + (1 - \alpha) F_{t-1}. \quad (12-1b)$$

When *smoothing constant* α is low, more weight is given to past data, and when it is high, more weight is given to recent data values. When α is equal to 0.9, then 99.99 per cent of the forecast value is determined by the four most recent demands. When α is as low as 0.1, only 34.39 per cent of the average is due to these last 4 periods and the smoothing effect is equivalent to a 19-period arithmetic moving average.

If α were assigned a value as high as 1, each forecast would reflect total adjustment to the recent data value and the forecast would simply be last period's actual value, that is, $F_t = 1.0D_{t-1}$. Since fluctuations are typically random, the value of α is generally kept in the range of 0.005 to 0.30 in order to 'smooth' the forecast.

The following table helps illustrate this concept. For example, when $\alpha = 0.5$, we can see that the new forecast is based on data value in the last three or four periods. When $\alpha = 0.1$, the forecast places little weight on recent value and takes a 19-period arithmetic moving average.

Smoothing Constant	Weight Assigned to				
	Most Recent Period (α)	2nd Most Recent Period $\alpha(1 - \alpha)$	3rd Most Recent Period $\alpha(1 - \alpha)^2$	4th Most Recent Period $\alpha(1 - \alpha)^3$	5th Most Recent Period $\alpha(1 - \alpha)^4$
$\alpha = 0.1$	0.1	0.09	0.081	0.073	0.066
$\alpha = 0.5$	0.5	0.25	0.125	0.063	0.031

Selecting the smoothing constant The exponential smoothing approach has been successfully applied by banks, manufacturing companies, wholesalers, and other organizations. The appropriate value of the exponential smoothing constant, α , however, can make the difference between an accurate and an inaccurate forecast. In picking a value for the smoothing constant, the objective is to obtain the most accurate forecast.

The correct α -value facilitates a reasonable reaction to a data value without incorporating too much random variation. An approximate value of α which is equivalent to an arithmetic moving average, in terms of degree of smoothing, can be estimated as: $\alpha = 2/(n + 1)$. The accuracy of a forecasting model can be determined by comparing the forecasted values with the actual or observed values.

Error The error of an individual forecast is defined as:

$$\text{Forecast error} = \text{Actual values} - \text{Forecasted values}$$

$$e_t = X_t - F_t$$

One measure of the overall forecast error for a model is the *mean absolute deviation (MAD)*. This is computed by taking the sum of the absolute values of the individual forecast errors and then dividing by number of periods n of data

$$\text{MAD} = \frac{\sum |\text{Forecast errors}|}{n}$$

where Standard deviation $\sigma \approx 1.25 \text{ MAD}$

Forecast: A projection or prediction of future values of a time-series.

The exponential smoothing method also facilitates continuous updating of the estimate of MAD. The current MAD_t is given by

$$\text{MAD}_t = \alpha | \text{Actual values} - \text{Forecasted values} | + (1 - \alpha) \text{MAD}_{t-1}$$

Higher values of smoothing constant α make the current MAD more responsive to current forecast errors

Example 12.7: A firm uses simple exponential smoothing with $\alpha = 0.1$ to forecast demand. The forecast for the week of February 1 was 500 units whereas actual demand turned out to be 450 units.

- (a) Forecast the demand for the week of February 8.
- (b) Assume the actual demand during the week of February 8 turned out to be 505 units. Forecast the demand for the week of February 15. Continue forecasting through March 15, assuming that subsequent demands were actually 516, 488, 467, 554, and 510 units.

Solution: Given $F_{t-1} = 500$, $D_{t-1} = 450$, and $\alpha = 0.1$

- (a) $F_t = F_{t-1} + \alpha(D_{t-1} - F_{t-1}) = 500 + 0.1(450 - 500) = 495$ units
- (b) Forecast of demand for the week of February 15 is shown in Table 12.5.

Table 12.5 Forecast of Demand

Week	Demand D_{t-1}	Old Forecast F_{t-1}	Forecast Error $(D_{t-1} - F_{t-1})$	Correction $\alpha(D_{t-1} - F_{t-1})$	New Forecast (F_t) $F_{t-1} + \alpha(D_{t-1} - F_{t-1})$
Feb. 1	450	500	-50	-5	495
8	505	495	10	1	496
15	516	496	20	2	498
22	488	498	-10	-1	497
Mar. 1	467	497	-30	-3	494
8	554	494	60	6	500
15	510	500	10	1	501

If no previous forecast value is known, the old forecast starting point may be estimated or taken to be an average of some preceding periods.

Example 12.8: A hospital has used a 9-month moving average forecasting method to predict drug and surgical inventory requirements. The actual demand for one item is shown in the table below. Using the previous moving average data, convert to an exponential smoothing forecast for month 33.

Month : 24	25	26	27	28	29	30	31	32
Demand : 78	65	90	71	80	101	84	60	73

Solution: The moving average of a 9-month period is given by

$$\text{Moving average} = \frac{\Sigma \text{Demand}(x)}{\text{Number of periods}} = \frac{78 + 65 + \dots + 73}{9} = 78$$

$$\text{Assume } F_{t-1} = 78. \text{ Therefore, estimated } \alpha = \frac{2}{n+1} = \frac{2}{9+1} = 0.2$$

$$\text{Thus, } F_t = F_{t-1} + \alpha(D_{t-1} - F_{t-1}) = 78 + 0.2(73 - 78) = 77 \text{ units}$$

Adjusted Exponential Smoothing

The simple exponential smoothing models is highly flexible because the smoothing effect can be increased or decreased by lowering or raising the value of α . However, if a trend exists in the data, the series will always lag behind the trend. Thus for an increasing trend the forecasts will be consistently low and for decreasing trends they will be consistently high. Simple exponential smoothing forecasts may be adjusted (F_t)_{adj} for trend effects by adding a trend smoothing factor β to the calculated forecast value F_t .

$$(F_t)_{\text{adj}} = F_t + \frac{1-\beta}{\beta} T_t$$

where $(F_t)_{\text{adj}}$ = trend-adjusted forecast

F_t = simple exponential smoothing forecast

β = smoothing constant for trend

T_t = exponentially smoothed trend factor

The value of the trend smoothing constant β , resembles the α constant in that a high β is more responsive to recent changes in trend. A low β gives less weight to the most recent trends and tends to smooth out the present trend. Values of β can be found by the trial-and-error approach, with the MAD used as a measure of comparison.

The value of the exponentially smoothed trend factor (T_t) is computed in a manner similar to that used in calculating the original forecast, and may be written as:

$$T_t = \beta(F_t - F_{t-1}) + (1 - \beta)T_{t-1}$$

where T_{t-1} = last period trend factor.

The trend factor T_t consists of a portion (β) of the trend evidenced from the current and previous forecast ($F_t - F_{t-1}$) with the remainder ($1 - \beta$) coming from the previous trend adjustment (T_{t-1}).

Simple exponential smoothing is often referred to as *first-order smoothing* and trend-adjusted smoothing is called *second-order* or *double smoothing*. Other advanced exponential smoothing models are also in use, including seasonal adjusted and triple smoothing.

Example 12.9: Develop an adjusted exponential forecast for the firm in Example 12.7. Assume the initial trend adjustment factor (T_{t-1}) is zero and $\beta = 0.1$.

Solution: Table 12.6 presents information needed to develop an adjusted exponential forecast.

Table 12.6

Week	D_{t-1}	F_{t-1}	F_t
Feb. 1	450	500	495
8	505	495	496
15	516	496	498
22	488	498	497
Mar. 1	467	497	494
8	554	494	500
15	510	500	501

The trend adjustment is an addition of a smoothing factor $\{(1 - \beta)/\beta\}T_t$ to the simple exponential forecast, so we need the previously calculated forecast values. Letting the first $T_{t-1} = 0$, we have

$$\begin{aligned} \text{Week 2/1: } T_t &= \beta(F_t - F_{t-1}) + (1 - \beta)T_{t-1} \\ &= 0.1(495 - 500) + (1 - 0.1)(0) = -0.50 \end{aligned}$$

$$\text{Adjusted forecast } (F_t)_{\text{adj}} = F_t + \frac{1-\beta}{\beta} T_t = 495 + \frac{1-0.1}{0.1}(-0.50) = 490.50$$

$$\text{Week 2/8: } T_t = 0.1(496 - 495) + 0.9(-0.50) = -0.35$$

$$\text{Adjusted forecast } (F_t)_{\text{adj}} = 496 + 9(-0.35) = 492.85$$

Putting the remainder of the calculations in table form, the trend-adjusted forecast for the week of March 15 is $(F_t)_{\text{adj}} = 501.44$ compared to the simple exponential forecast of $F_t = 500$, which is not a large difference.

Self-Practice Problems 12A

- 12.1** The owner of a small company manufactures a product. Since he started the company, the number of units of the product he has sold is represented by the following time series:

Year :	1995	1996	1997	1998	1999	2000	2001
Units sold	100	120	95	105	108	102	112

Find the trend line that describes the trend by using the method of semi-averages.

- 12.2** Fit a trend line to the following data by the freehand method:

Year	Production of Steel (million tonnes)	Year	Production of Steel (million tonnes)
1995	20	2000	25
1996	22	2001	23
1997	24	2002	26
1998	21	2003	25
1999	23		

- 12.3** A State Govt. is studying the number of traffic fatalities in the state resulting from drunken driving for each of the last 12 months

Month	Accidents
1	280
2	300
3	280
4	280
5	270
6	240
7	230
8	230
9	220
10	200
11	210
12	200

Find the trend line that describes the trend by using the method of semi-averages.

- 12.4** Calculate the three-month moving averages from the following data:

Jan.	Feb.	March	April	May	June
57	65	63	72	69	78
July	Aug.	Sept.	Oct.	Nov.	Dec.
82	81	90	92	95	97

[Osmania Univ., BCom, 1996]

- 12.5** Gross revenue data (Rs in million) for a Travel Agency for a 11-year period is as follows:

Year	Revenue
1995	3
1996	6
1997	10
1998	8
1999	7
2000	12
2001	14
2002	14
2003	18
2004	19

Calculate a 3-year moving average for the revenue earned.

- 12.6** The owner of small manufacturing company has been concerned about the increase in manufacturing costs over the past 10 years. The following data provide a time series of the cost per unit for the company's leading product over the past 10 years.

Year	Cost per Unit	Year	Cost per Unit
1995	332	2000	405
1996	317	2001	410
1997	357	2002	427
1998	392	2003	405
1999	402	2004	438

Calculate a 5-year moving average for the unit cost of the product.

- 12.7** The following data provide a time series of the number of Commercial and Industrial units failures during the period 1989–2004.

Year	No. of Failures	Year	No. of Failures
1989	23	1997	9
1990	26	1998	13
1991	28	1999	11
1992	32	2000	14
1993	20	2001	12
1994	12	2002	9
1995	12	2003	3
1996	10	2004	1

Calculate a 5-year and 7-year moving average for the number of units failure.

- 12.8** Estimate the trend values using the data given by taking a four-year moving average :

Year	Value	Year	Value
1990	12	1997	100
1991	25	1998	82
1992	39	1999	65
1993	54	2000	49
1994	70	2001	34
1995	87	2002	20
1996	105	2003	7

[Madras Univ., MCom, 1998]

12.9 In January, a city hotel predicted a February demand for 142 room occupancy. Actual February demand was 153 rooms. Using a smoothing constant of $\alpha = 0.20$, forecast the March demand using the exponential smoothing model.

12.10 A shoe manufacturer, using exponential smoothing with $\alpha = 0.1$, has developed a January trend forecast of 400 units for a ladies' shoe. This brand has seasonal indexes of 0.80, 0.90, and 1.20 respectively for the first three months of the year. Assuming actual sales were 344 units in January and 414 units in February, what would be the seasonalized March forecast?

12.11 A food processor uses exponential smoothing (with $\alpha = 0.10$) to forecast next month's demand. Past (actual) demand in units and the simple exponential forecasts up to month 51 are shown in the following table

Month	Actual Demand	Old Forecast
43	105	100.00
44	106	100.50
45	110	101.05
46	110	101.95
47	114	102.46
48	121	103.61
49	130	105.35
50	128	107.82
51	137	109.84

- (a) Using simple exponential smoothing, forecast the demand for month 52.
 (b) Suppose a firm wishes to start including a trend-adjustment factor of $\beta = 0.60$. If it assumes an initial trend adjustment of zero ($T_t = 0$) in month 50, what would be the value of $(F_t)_{\text{adj}}$ for month 52?

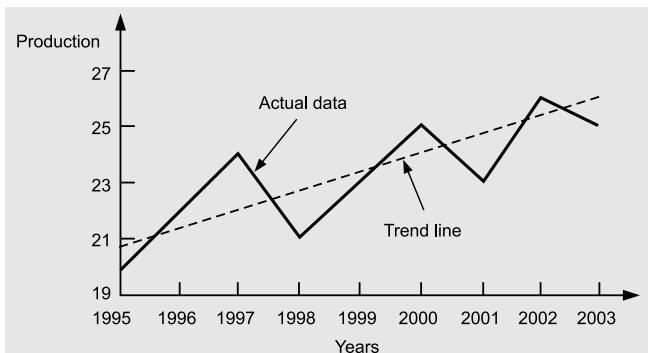
Hints and Answers

12.1

Year (y)	Units Sold (x)
1995	100
1996	120
1997	95
1998	105
1999	108
2000	102
2001	112

Trend line $y = 105 + 107.33x$.

12.2



12.3

Month	Accidents
1	280
2	300
3	280
4	280
5	270
6	240
7	230
8	230
9	220
10	200
11	210
12	200

Average of first 6 months, $a = 1650/6 = 275$

Average of last 6 months, $b = 1290/6 = 215$

Trend line $y = 275 + 215x$.

12.4

Month	Values	3-month Total	3-month Moving Average
Jan.	57	—	—
Feb.	65	185	$185/3 = 61.67$
March	63	200	$200/3 = 66.67$
April	72	204	$204/3 = 68.00$
May	69	219	73.00
June	78	229	76.33
July	82	241	80.33
Aug.	81	253	84.33
Sept.	90	263	87.67
Oct.	92	277	92.38
Nov.	95	284	94.67
Dec.	97	—	—

12.5

Year	Revenue	3-year Moving Total	3-year Moving Average
1995	3	—	—
1996	6	19	$19/3 = 6.33$
1997	10	24	$24/3 = 8.00$
1998	8	21	$21/3 = 7.00$
1999	7	25	8.33
2000	12	32	10.66
2001	14	34	11.33
2002	14	46	15.33
2003	18	51	17.00
2004	19	—	—

12.6

Year	Per Unit Cost	5-year Moving Total	5-year Moving Average
1995	332	—	—
1996	317	—	—
1997	357	1800	$1800/5 = 360.0$
1998	392	1873	$1873/5 = 374.6$
1999	402	1966	$1966/5 = 393.2$
2000	405	2036	407.2
2001	410	2049	409.8
2002	427	2085	417.0
2003	405	—	—
2004	438	—	—

12.7

Year	Number of Failures	5-year Moving Total	5-year Moving Average	7-year Moving Total	7-year Moving Average
1989	23	—	—	—	—
1990	26	—	—	—	—
1991	28	→ 129	25.8	—	—
1992	32	→ 118	23.6	153	21.9
1993	20	104	20.8	140	20.0
1994	12	86	17.2	123	17.6
1995	12	63	12.6	108	15.4
1996	10	56	11.2	87	12.4
1997	9	55	11.0	81	11.6
1998	13	57	11.4	81	11.6
1999	11	59	11.8	78	11.1
2000	14	59	11.8	71	10.1
2001	12	69	9.8	63	5.0
2002	9	39	7.9	—	—
2003	3	—	—	—	—
2004	1	—	—	—	—

12.8

Year	Value	4-year Centred Total	4-year Moving Average	Moving Average
1990	12	—	—	—
1991	25	—	—	—
1992	39	→ 130	$130/4=32.5$	$(32.5 + 47)/2 = 39.75$
1993	54	→ 188	$188/4=47.0$	$(47 + 62.5)/2 = 54.75$
1994	70	→ 250	$250/4=62.5$	70.75
1995	87	316	79.0	84.75
1996	105	362	90.5	92.00
1997	100	374	93.5	90.75
1998	82	352	88.0	81.00
1999	65	296	74.0	65.75
2000	49	230	57.5	49.75
		168	42.0	
2001	34	110	27.5	34.75
2002	20	—	—	—
2003	7	—	—	—

12.9 New forecast (March demand)

$$\begin{aligned} &= F_{t-1} + \alpha(D_{t-1} - F_{t-1}) \\ &= 142 + 0.20(153 - 142) = 144.20 = 144 \end{aligned}$$

rooms

- 12.10** (a) Deseasonalized actual January demand
 $= 344/0.80 = 430$ units

(b) Compute the deseasonalized forecast

$$\begin{aligned} F_t &= F_{t-1} + \alpha(D_{t-1} - F_{t-1}) \\ &= 400 + 0.1(430 - 400) = 403 \end{aligned}$$

- 12.11** In this problem the smoothing constant for the original data ($\alpha = 0.10$) differs from the smoothing constant for the trend $\beta = 0.60$.

$$\begin{aligned} (a) F_t &= F_{t-1} + \alpha(D_{t-1} - F_{t-1}) \\ &= 109.84 + 0.1(137.00 - 109.84) = 112.56 \end{aligned}$$

(b) Forecast for month 51 :

$$(F_t)_{\text{adj}} = F_t + \frac{1-\beta}{\beta} T_t$$

$$\begin{aligned} \text{where } T_t &= \beta(F_t - F_{t-1}) + (1-\beta)T_{t-1} \\ &= 0.6(109.84 - 107.82) + (1-0.6)0 \\ &= 1.21 \end{aligned}$$

$$\begin{aligned} (F_t)_{\text{adj}} &= 109.84 + \left(\frac{1-0.6}{0.6}\right)(1.21) \\ &= 110.65 \end{aligned} \quad (1.21)$$

Forecast for month 52 :

$$\begin{aligned} T_t &= \beta(F_t - F_{t-1}) + (1-\beta)T_{t-1} \\ &= 0.6(112.56 - 109.84) + (1-0.6)(1.21) \\ &= 2.12 \end{aligned} \quad (1.21)$$

$$\begin{aligned} (F_t)_{\text{adj}} &= F_t + \frac{1-\beta}{\beta} T_t \\ &= 112.56 + \left(\frac{1-0.6}{0.6}\right)(2.12) = 113.98 \end{aligned} \quad (2.12)$$

12.9 TREND PROJECTION METHODS

A *trend* is the long-run general direction (upward, downward or constant) of a business climate over a period of several years. It is best represented by a straight line.

The trend projection method fits a trend line to a time series data and then projects medium-to-long-range forecasts. Several possible trend fits can be explored (such as exponential and quadratic), depending upon movement of time-series data. In this section, we will discuss linear, quadratic and exponential trend models. Since seasonal effects can compound trend analysis, it is assumed that no seasonal effects occur in the data or are removed before establishing the trend.

Reasons to study trend: A few reasons to study trends are as follows:

1. The study of trend helps to describe the long-run general direction (upward, downward, constant) of a business climate over a period of several years.
2. The study allows us to use trends as an aid in making intermediate and long-range forecasting projections in the future.
3. The study of trends help to isolate and then eliminate its influencing effects on the time-series model.

12.9.1 Linear Trend Model

The *method of least squares* from regression analysis is used to find the *trend line of best fit* to a time series data. The regression trend line (y) is defined by the following equation.

$$\hat{y} = a + bx$$

where \hat{y} = predicted value of the dependent variable

a = y -axis intercept,

b = slope of the regression line (or the rate of change in y for a given change in x),

x = independent variable (which is *time* in this case).

The trend line of best fit has the properties that (i) the summation of all vertical deviations about it is zero, that is, $\sum(y - \hat{y}) = 0$, (ii) the summation of all vertical deviations squared is a minimum, that is, $\sum(y - \hat{y})^2$ is least, and (iii) the line goes through the mean values of variables x and y . For linear equations, it is found by the simultaneous solution for a and b of the two normal equations:

$$\Sigma y = na + b\Sigma x \quad \text{and} \quad \Sigma xy = a\Sigma x + b\Sigma x^2$$

where the data can be coded so that $\Sigma x = 0$, two terms in these equations drop out and we have

$$\Sigma y = na \quad \text{and} \quad \Sigma xy = b \Sigma x^2$$

Coding is easily done with time-series data. For coding the data, we choose the centre of the time period as $x = 0$ and have an equal number of plus and minus periods on each side of the trend line which sum to zero.

Alternately, we can also find the values of constants a and b for any regression line as:

$$b = \frac{\Sigma xy - n \bar{x} \bar{y}}{\Sigma x^2 - n(\bar{x})^2} \quad \text{and} \quad a = \bar{y} - b \bar{x}$$

Example 12.10: Below are given the figures of production (in thousand quintals) of a sugar factory:

Year	:	1995	1996	1997	1998	1999	2000	2001
Production	:	80	90	92	83	94	99	92

(a) Fit a straight line trend to these figures

(b) Plot these figures on a graph and show the trend line.

(c) Estimate the production in 2004. [Bangalore Univ., BCom, 1998]

Solution: (a) Using normal equations and the sugar production data we can compute constants a and b as shown in Table 12.7:

Table 12.7 Calculation for Least Squares Equation

Year	Time Period (x)	Production (x)	x^2	xy	Trend Values \hat{y}
1995	1	80	1	80	84
1996	2	90	4	180	86
1997	3	92	9	276	88
1998	4	83	16	332	90
1999	5	94	25	470	92
2000	6	99	36	594	94
2001	7	92	49	644	96
	28	630	140	2576	

$$\bar{x} = \frac{\Sigma x}{n} = \frac{28}{7} = 4, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{630}{7} = 90$$

$$b = \frac{\Sigma xy - n \bar{x} \bar{y}}{\Sigma x^2 - n(\bar{x})^2} = \frac{2576 - 7(4)(90)}{140 - 7(4)^2} = \frac{56}{28} = 2$$

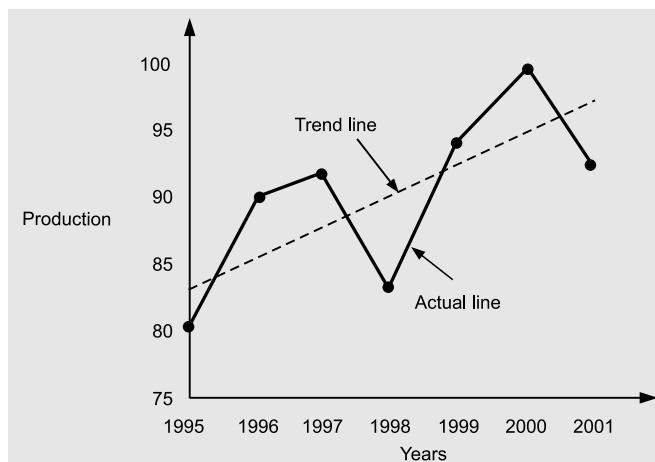
$$a = \bar{y} - b \bar{x} = 90 - 2(4) = 82$$

Therefore, linear trend component for the production of sugar is:

$$\hat{y} = a + bx = 82 + 2x$$

The slope $b = 2$ indicates that over the past 7 years, the production of sugar had an average growth of about 2 thousand quintals per year.

Figure 12.7
Linear Trend for Production of Sugar



(b) Plotting points on the graph paper, we get an actual graph representing production of sugar over the past 7 years. Join the point $a = 82$ and $b = 2$ (corresponds to 1996) on the graph we get a trend line as shown in Fig. 12.7.

(c) The production of sugar for year 2004 will be

$$\hat{y} = 82 + 2(10) = 102 \text{ thousand quintals}$$

Example 12.11: The following table relates to the tourist arrivals (in millions) during 1994 to 2000 in India:

Year	:	1994	1995	1996	1997	1998	1999	2000
Tourists arrivals :		18	20	23	25	24	28	30

Fit a straight line trend by the method of least squares and estimate the number of tourists that would arrive in the year 2004. [Kurukshetra Univ., MTM., 1997]

Solution: Using normal equations and the tourists arrival data we can compute constants a and b as shown in Table 12.8:

Table 12.8 Calculations for Least Squares Equation

Year	Time Scale (x)	Tourist Arrivals (y)	xy	x^2
1994	-3	18	-54	9
1995	-2	20	-40	4
1996	-1	23	-23	1
1997	0	25	0	0
1998	1	24	24	1
1999	2	28	56	4
2000	3	30	90	9
		168	53	28

$$\bar{x} = \frac{\sum x}{n} = 0, \bar{y} = \frac{\sum y}{n} = \frac{168}{7} = 24$$

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2} = \frac{53}{28} = 1.893;$$

$$a = \bar{y} - b\bar{x} = 24 - 1.893(0) = 24$$

Therefore, the linear trend component for arrival of tourists is

$$\hat{y} = a + bx = 24 + 1.893x$$

The estimated number of tourists that would arrive in the year 2004 are:

$$\hat{y} = 24 + 1.893(7) = 37.251 \text{ million (measured from 1997 = origin)}$$

12.9.2 Quadratic Trend Model

The quadratic relationship for estimating the value of a dependent variable y from an independent variable x might take the form

$$\hat{y} = a + bx + cx^2$$

This trend line is also called the *parabola*.

For a non-linear equation $y = a + bx + cx^2$, the values of constants a , b , and c can be determined by solving three normal equations

$$\begin{aligned} \Sigma y &= na + b\Sigma x + c\Sigma x^2 \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \\ \Sigma x^2y &= a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \end{aligned}$$

When the data can be coded so that $\Sigma x = 0$ and $\Sigma x^3 = 0$, two terms in the above expressions drop out and we have

$$\begin{aligned} \Sigma y &= na + c\Sigma x^2 \\ \Sigma xy &= b\Sigma x^2 \\ \Sigma x^2y &= a\Sigma x^2 + c\Sigma x^4 \end{aligned}$$

To find the exact estimated value of the variable y , the values of constants a , b , and c need to be calculated. The values of these constants can be calculated by using the following shortest method:

$$a = \frac{\Sigma y - c \Sigma x^2}{n}; \quad b = \frac{\Sigma x y}{\Sigma x^2} \text{ and } c = \frac{n \Sigma x^2 y - \Sigma x^2 \Sigma y}{n \Sigma x^4 - (\Sigma x^2)^2}$$

Example 12.12: The prices of a commodity during 1998–2003 are given below. Fit a parabola to these data. Estimate the price of the commodity for the year 2004.

Year	Price	Year	Price
1998	100	2001	140
1999	107	2002	181
2000	128	2003	192

Also plot the actual and trend values on a graph.

Solution: To fit a quadratic equation $\hat{y} = a + bx + cx^2$, the calculations to determine the values of constants a , b , and c are shown in Table 12.9.

Table 12.9 Calculations for Parabola Trend Line

Year	Time Scale (x)	Price (y)	x^2	x^3	x^4	xy	$x^2 y$	Trend Values (\hat{y})
1998	-2	100	4	-8	16	-200	400	97.72
1999	-1	107	1	-1	1	-107	107	110.34
2000	0	128	0	0	0	0	0	126.68
2001	1	140	1	1	1	140	140	146.50
2002	2	181	4	8	16	362	724	169.88
2003	3	192	9	27	81	576	1728	196.82
	3	848	19	27	115	771	3099	847.94

$$(i) \quad \Sigma y = n a + b \Sigma x + c \Sigma x^2 \quad \text{or} \quad 848 = 6a + 3b + 19c$$

$$(ii) \quad \Sigma xy = a \Sigma x + b \Sigma x^2 + c \Sigma x^3 \quad \text{or} \quad 771 = 3a + 19b + 27c$$

$$(iii) \quad \Sigma x^2 y = a \Sigma x^2 + b \Sigma x^3 + c \Sigma x^4 \quad \text{or} \quad 3099 = 19a + 27b + 115c$$

Eliminating a from eqns. (i) and (ii), we get

$$(iv) \quad 694 = 35b + 35c$$

Eliminating a from eqns. (ii) and (iii), we get

$$(v) \quad 5352 = 280b + 168c$$

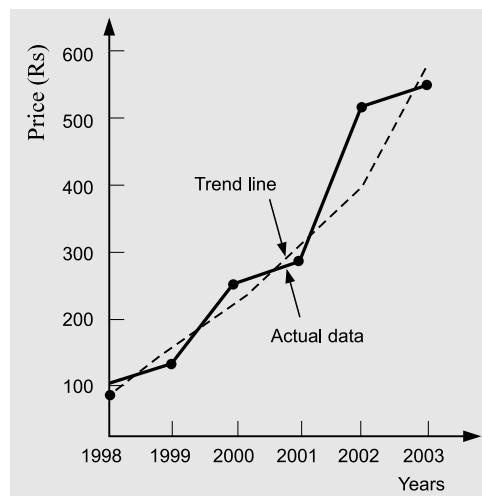
Solving eqns. (iv) and (v) for b and c we get $b = 18.04$ and $c = 1.78$. Substituting values of b and c in eqn. (i), we get $a = 126.68$.

Hence, the required non-linear trend line becomes

$$y = 126.68 + 18.04x + 1.78x^2$$

Several trend values as shown in Table 12.9 can be obtained by putting $x = -2, -1, 0, 1, 2$, and 3 in the trend line. The trend values are plotted on a graph paper. The graph is shown in Fig. 12.8.

Figure 12.8
Trend Line for Price of Commodity



12.9.3 Exponential Trend Model

When the given values of dependent variable y form approximately a geometric progression while the corresponding independent variable x values form an arithmetic progression, the relationship between variables x and y is given by an exponential function, and the best fitting curve is said to describe the *exponential trend*. Data from the fields of biology, banking, and economics frequently exhibit such a trend. For example, growth of bacteria, money accumulating at compound interest, sales or earnings over a short period, and so on, follow exponential growth.

The characteristic property of this law is that the rate of growth, that is, the rate of change of y with respect to x is proportional to the values of the function. The following function has this property.

$$y = a b^{cx}, a > 0$$

The letter b is a fixed constant, usually either 10 or e , where a is a constant to be determined from the data.

To assume that the law of growth will continue is usually unwarranted, so only short range predictions can be made with any considerable degree of reliability.

If we take logarithms (with base 10) of both sides of the above equation, we obtain

$$\log y = \log a + (c \log b) x$$

For $b = 10$, $\log b = 1$, but for $b = e$, $\log b = 0.4343$ (approx.). In either case, this equation is of the form

$$y' = c + dx \quad (12-2)$$

where $y' = \log y$, $c = \log a$, and $d = c \log b$.

Equation (12-2) represents a straight line. A method of fitting an exponential trend line to a set of observed values of y is to fit a straight trend line to the logarithms of the y -values.

In order to find out the values of constants a and b in the exponential function, the two normal equations to be solved are

$$\Sigma \log y = n \log a + \log b \Sigma x$$

$$\Sigma x \log y = \log a \Sigma x + \log b \Sigma x^2$$

When the data is coded so that $\Sigma x = 0$, the two normal equations become

$$\Sigma \log y = n \log a \quad \text{or} \quad \log a = \frac{1}{n} \Sigma \log y$$

$$\text{and} \quad \Sigma x \log y = \log b \Sigma x^2 \quad \text{or} \quad \log b = \frac{\Sigma x \log y}{\Sigma x^2}$$

Coding is easily done with time-series data by simply designating the center of the time period as $x = 0$, and have equal number of plus and minus period on each side which sum to zero.

Example 12.13: The sales (Rs in million) of a company for the years 1995 to 1999 are:

Year	1997	1998	1999	2000	2001
Sales	1.6	4.5	13.8	40.2	125.0

Find the exponential trend for the given data and estimate the sales for 2004.

Solution: The computational time can be reduced by coding the data. For this consider $u = x - 3$. The necessary computations are shown in Table 12.10.

Table 12.10 Calculation for Least Squares Equation

Year	Time Period x	$u = x - 3$	u^2	Sales y	$\log y$	$u \log y$
1997	1	-2	4	1.60	0.2041	-0.4082
1998	2	-1	1	4.50	0.6532	-0.6532
1999	3	0	0	13.80	1.1390	0
2000	4	1	1	40.20	1.6042	1.6042
2001	5	2	4	125.00	2.0969	4.1938
			10		5.6983	4.7366

$$\log a = \frac{1}{n} \sum \log y = \frac{1}{5} (5.6983) = 1.1397$$

$$\log b = \frac{\sum u \log y}{\sum u^2} = \frac{4.7366}{10} = 0.4737$$

Therefore $\log y = \log a + (x + 3) \log b = 1.1397 + 0.4737x$

For sales during 2004, $x = 3$, and we obtain

$$\log y = 1.1397 + 0.4737 (3) = 2.5608$$

or

$$y = \text{antilog}(2.5608) = 363.80$$

12.9.4 Changing the Origin and Scale of Equations

When a moving average or trend value is calculated it is assumed to be centred in the middle of the month (fifteenth day) or the year (July 1). Similarly, the forecast value is assumed to be centred in the middle of the future period. However, the reference point (origin) can be shifted, or the units of variables x and y are changed to monthly or quarterly values if desired. The procedure is as follows:

- (i) Shift the origin, simply by adding or subtracting the desired number of periods from independent variable x in the original forecasting equation.
- (ii) Change the time units from annual values to monthly values by dividing independent variable x by 12.
- (iii) Change the y units from annual to monthly values, the entire right-hand side of the equation must be divided by 12.

Example 12.14: The following forecasting equation has been derived by a least-squares method:

$$\hat{y} = 10.27 + 1.65x \quad (\text{Base year: 1997; } x = \text{years; } y = \text{tonnes/year})$$

Rewrite the equation by

- (a) shifting the origin to 2002.
- (b) expressing x units in months, retaining y in tonnes/year.
- (c) expressing x units in months and y in tonnes/month.

Solution: (a) Shifting of origin can be done by adding the desired number of period 5 (1997 to 2002) to x in the given equation. That is

$$\hat{y} = 10.27 + 1.65(x + 5) = 18.52 + 1.65x$$

where 2002 = 0, x = years, y = tonnes/year.

(b) Expressing x units in months

$$\hat{y} = 10.27 + \frac{1.65x}{12} = 10.27 + 0.14x$$

where July 1, 1997 = 0, x = months, y = tonnes/year.

(c) Expressing y in tonnes/month, retaining x in months

$$\hat{y} = \frac{1}{12}(10.27 + 0.14x) = 0.86 + 0.01x$$

where July 1, 1997 = 0, x = months, y = tonnes/month.

Remarks

1. If both x and y are to be expressed in months together, then divide constant ' a ' by 12 and constant ' b ' by 24. It is because data are sums of 12 months. Thus monthly trend equation becomes

$$\text{Linear trend : } \hat{y} = \frac{a}{12} + \frac{b}{24}x$$

$$\text{Parabolic trend : } \hat{y} = \frac{a}{12} + \frac{b}{144}x + \frac{c}{1728}x^2$$

But if data are given as monthly averages per year, then value of 'a' remains unchanged, 'b' is divided by 12 and 'c' by 144.

2. The annual trend equation can be reduced to quarterly trend equation as:

$$\hat{y} = \frac{a}{4} + \frac{b}{4 \times 12} x = \frac{a}{4} + \frac{b}{48} x$$

Self-Practice Problems 12B

- 12.12** The general manager of a building materials production plant feels that the demand for plasterboard shipments may be related to the number of construction permits issued in the country during the previous quarter. The manager has collected the data shown in the table.

Construction Permits	Plasterboard Shipments
15	6
9	4
40	16
20	6
25	13
25	9
15	10
35	16

- (a) Use the normal equations to derive a regression forecasting equation.
- (b) Determine a point estimate for plasterboard shipments when the number of construction permits is 30.

- 12.13** A company that manufactures steel observed the production of steel (in metric tonnes) represented by the time-series:

Year : 1996 1997 1998 1999 2000 2001 2002
Production
of steel : 60 72 75 65 80 85 95

- (a) Find the linear equation that describes the trend in the production of steel by the company.
- (b) Estimate the production of steel in 2003.

- 12.14** Fit a straight line trend by the method of least squares to the following data. Assuming that the same rate of change continues, what would be the predicted earning (Rs in lakh) for the year 2004?

Year : 1995 1996 1997 1998 1999 2000 2001 2002
Earnings : 38 40 65 72 69 60 87 95

[Agra Univ., BCom 1996; MD Univ., BCom, 1998]

- 12.15** The sales (Rs in lakh) of a company for the years 1990 to 1996 are given below:

Year : 1998 1999 2000 2001 2002 2003 2004
Sales : 32 47 65 88 132 190 275

Find trend values by using the equation $y_c = a b^x$ and estimate the value for 2005.

[Delhi Univ., BCom, 1996]

- 12.16** A company that specializes in the production of petrol filters has recorded the following production (in 1000 units) over the last 7 years.

Years : 1995 96 97 98 99 00 01
Production : 42 49 62 75 92 122 158

- (a) Develop a second-degree estimating equation that best describes these data.
- (b) Estimate the production in 2005.

- 12.17** In 1996 a firm began downsizing in order to reduce its costs. One of the results of these cost cutting measures has been a decline in the percentage of private industry jobs that are managerial. The following data show the percentage of females who are managers from 1996 to 2003.

Years : 1996 97 98 99 00 01 02 03
Percentage : 6.7 5.3 4.3 6.1 5.6 7.9 5.8 6.1

- (a) Develop a linear trend line for this time series through 2001 only.
- (b) Use this trend to estimate the percentage of females who are managers in 2004.

- 12.18** A company develops, markets, manufactures, and sells integrated wide-area network access products. The following are annual sales (Rs in million) data from 1998 to 2004.

Year : 1998 1999 2000 2001 2002 2003 2004
Sales : 16 17 25 28 32 43 50

- (a) Develop the second-degree estimating equation that best describes these data.
- (b) Use the trend equation to forecast sales for 2005.

Hints and Answers

12.12 (a)

x	y	xy	x^2	y^2
15	6	90	225	36
9	4	36	81	16
40	16	640	1,600	256
20	6	120	400	36
25	13	325	625	169
25	9	225	625	81
15	10	150	225	100
35	16	560	1,225	256
184	80	2,146	5,006	950

n = 8 pairs of observations;

$$\bar{x} = 184/8 = 23; \bar{y} = 80/8 = 10$$

$$\Sigma y = na + b\Sigma x \quad \text{or} \quad 80 = 8a + 184b$$

$$\Sigma xy = \Sigma x + b\Sigma x^2 \quad \text{or} \quad 2,146 = 184a + 5,006b$$

After solving equations we get $a = 0.91$ and $b = 0.395$.

Therefore the equation is: $\hat{y} = 0.91 + 0.395x$

(b) For $x = 30$, we have $\hat{y} = 0.91 + 0.395(30) = 13$ shipments (approx.)

12.13 $a = \Sigma y/n = 532/7 = 76$; $b = \Sigma xy/\Sigma x^2 = 136/28 = 4.857$

(a) Trend line $\hat{y} = a + bx = 76 + 4.857x$

(b) For 2003, $x = 4$, $\hat{y} = 76 + 4.857(4) = 95.428$ metric tonnes.

12.14 $a = \Sigma y/n = 526/8 = 65.75$;

$$b = \Sigma xy/\Sigma x^2 = 616/168 = 3.667$$

Trend line : $\hat{y} = a + bx = 65.75 + 3.667x$

For 2004, $x = 11$; $\hat{y} = 65.75 + 3.667(11)$
= Rs 106.087 lakh.

12.15 $\log a = \frac{1}{n} \sum \log y = \frac{1}{7}(13.7926) = 1.9704$

$$\log b = \frac{\sum x \log y}{\sum x^2} = \frac{4.3237}{28} = 0.154$$

Thus $\log y = \log a + x \log b = 1.9704 + 0.154x$

For 2005, $x = 4$; $\log y = 1.9704 + 0.154(4)$
= 2.5864

$y = \text{Antilog}(2.5864) = \text{Rs } 385.9$ lakh.

12.16

Year	Period	Deviation from 1998 (x)	x^2	x^4	y	xy	x^2y
1995	1	-3	9	81	42	-126	378
1996	2	-2	4	16	49	-98	196
1997	3	-1	1	1	62	-62	62
1998	4	0	0	0	75	0	0
1999	5	1	1	1	92	+92	92
2000	6	2	4	16	122	+244	488
2001	7	3	9	81	158	+474	1422
		0	28	196	600	524	2638

(a) Solving the equations

$$\Sigma y = na + c\Sigma x^2 \quad \text{or} \quad 600 = 7a + 28c$$

$$\Sigma x^2y = a\Sigma x^2 + c\Sigma x^4 \quad \text{or} \quad 2638 = 28a + 196c$$

$$\Sigma xy = b\Sigma x^2 \quad \text{or} \quad 524 = 28b$$

We get $a = 80.05$, $b = 18.71$ and $c = -1.417$

$$\text{Hence } \hat{y} = a + bx + cx^2 = 80.05 + 18.71x - 1.417x^2$$

(b) For 2005, $x = 8$; $\hat{y} = 80.05 + 18.71(8) - 1.417(8)^2$
= Rs 139.042 thousand.

12.17

Year	Time Period	Deviation from 2001	Percentage of Females		xy	x^2
			x	y		
1996	1	-5	6.7	-33.5	25	
1997	2	-4	5.3	-21.2	16	
1998	3	-3	4.3	-12.9	9	
1999	4	-2	6.1	-12.2	4	
2000	5	-1	5.6	-6.6	1	
2001	6	0	7.9	0	0	
2002	7	1	5.8	5.8	1	
2003	8	2	6.1	12.2	4	
		-12	47.8	-68.4	60	

(a) Solving the equations

$$\Sigma y = na + b\Sigma x \quad \text{or} \quad 47.8 = 8a - 12b$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \text{or} \quad -67.4 = -12a + 60b$$

We get $a = 6.28$ and $b = 0.102$

$$\text{Hence } \hat{y} = a + bx = 6.128 + 0.102x$$

(b) For 2004, $x = 3$; $\hat{y} = 6.128 + 0.102(3)$
= 6.434 per cent.

12.18

Year	Time Period	Deviation from 2001 (x)	Sales		xy	x^2	x^4	x^2y
			x	y				
1998	1	-3	16	-48	9	81	144	
1999	2	-2	17	-34	4	16	68	
2000	3	-1	25	-25	1	1	25	
2001	4	0	28	0	0	0	0	
2002	5	1	32	32	1	1	32	
2003	6	2	43	86	4	16	172	
2004	7	3	50	150	9	81	450	
		0	211	161	28	196	891	

(a) Solving the equations

$$\Sigma y = na + c\Sigma x^2 \quad \text{or} \quad 211 = 7a + 28c$$

$$\Sigma x^2y = a\Sigma x^2 + c\Sigma x^4 \quad \text{or} \quad 891 = 28a + 196c$$

$$\Sigma xy = b\Sigma x^2 \quad \text{or} \quad 161 = 28b$$

We get $a = 27.904$, $b = 5.75$ and $c = 0.559$

$$\hat{y} = a + bx + cx^2 = 27.904 + 5.75x + 0.559x^2$$

For 2005, $x = 4$; $\hat{y} = 27.904 + 5.75(4) + 0.559(4)^2$
= 59.848

12.10 MEASUREMENT OF SEASONAL EFFECTS

As mentioned earlier that time-series data consists of four components: trend, cyclical effects, seasonal effects and irregular fluctuations. In this section, we will discuss techniques for identifying seasonal effects in a time-series data. Seasonal effect is defined as the repetitive and predictable pattern of data behaviour in a time-series around the trend line during particular time intervals of the year. In order to measure (or detect) the seasonal effect, time period must be less than one year such as days, weeks, months, or quarters.

Seasonal effects arises as the result of natural changes in the seasons during the year or may result due to habits, customs, or festivals that occur at the same time year after year.

We have three main reasons to study seasonal effects:

- (i) The description of the seasonal effect provides a better understanding of the impact this component has upon a particular time-series.
- (ii) Once the seasonal pattern that exists is established, seasonal effect can be eliminated from the time-series in order to observe the effect of the other components, such as cyclical and irregular components. Elimination of seasonal effect from the series is referred to as **deseasonalizing** or **seasonal adjusting** of the data.
- (iii) Trend analysis may be adequate for long-range forecast, but for short-run predictions, knowledge of seasonal effects on time-series data is essential for projection of past pattern into the future.

Remarks:

1. In an additive time-series model, we can estimate the seasonal component as:

$$S = Y - (T + C + I)$$

In the absence of C and I, we have $S = Y - T$. That is, the seasonal component is the difference between actual data values in series and the trend values.

2. One of the technique for isolating the effects of seasonality is decomposition. The process of decomposition begins by determining T.C for each and dividing the time-series data (T.C.S.I) by T.C. The resulting expression contains seasonal effects along with irregular fluctuations

$$\frac{T.C.S.I}{T.C} = S.I.$$

A method for eliminating irregular fluctuations can be applied, leaving only the seasonal effects as shown below.

$$\text{Seasonal effect} = \frac{T.S.C.I}{T.C.I} = \frac{Y}{T.C.I} \times 100\%$$

3. The process of eliminating the effects of seasonality from a time-series data is referred to as **de-seasonalization** or **seasonal adjustment**. The data can be deseasonalized by dividing the actual values Y by final adjusted seasonal effects, and is expressed as:

$$\frac{Y}{S} = \frac{T.S.C.I}{S} = T.C.I \times 100\% \quad \leftarrow \text{Multiplicative}$$

$$Y - S = (T + S + C + I) - S = T + C + I \quad \leftarrow \text{Additive Model}$$

Each adjusted seasonal index measures the average magnitude of seasonal influence on the actual values of the time series for a given period within a year. By subtracting the base index of 100 (which represents the T and C components) from each seasonal index, the extent of the influence of seasonal force can be measured.

Deseasonalization: A statistical process used to remove the effect of seasonality from a time-series by dividing each original series observation by the corresponding seasonal index.

12.10.1 Seasonal Index

Seasonal effects are measured in terms of an index, called *seasonal index*, attached to each period of the time series within a year. Hence, if monthly data are considered, there are 12 separate seasonal indexes, one for each month. Similarly for quarterly data, there are 4 separate indexes. A *seasonal index* is an average that indicates the percentage deviation of actual values of the time series from a base value which excludes the short-term seasonal influences. The base time series value represents the trend/cyclical influences only.

The following four methods are used to construct seasonal indexes to measure seasonal effects in the time-series data:

- (i) Method of simple averages
- (ii) Ratio-to-trend method
- (iii) Ratio-to-moving average method
- (iv) Link relatives method.

12.10.2 Method of Simple Averages

This method is also called *average percentage method* because this method expresses the data of each month or quarter as a percentage of the average of the year. The steps of the method are summarized below:

- (i) Average the unadjusted data by years and months (or quarters if quarterly data are given).
- (ii) Add the figures of each month and obtain the averages by dividing the monthly totals by the number of years. Let the averages for 12 months be denoted by \bar{x}_1 , \bar{x}_2 , ..., \bar{x}_{12} .
- (iii) Obtain an average of monthly averages by dividing the total of monthly averages by 12. That is

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \dots + \bar{x}_{12}}{12}$$

- (iv) Compute seasonal indexes for different months by expressing monthly averages as percentages of the grand average $\bar{\bar{x}}$ as follows:

$$\begin{aligned} \text{Seasonal index for month } i &= \frac{\text{Monthly average for month } i}{\text{Average of monthly averages}} \times 100 \\ &= \frac{\bar{x}_i}{\bar{\bar{x}}} \times 100 \quad (i = 1, 2, \dots, 12) \end{aligned}$$

It is important to note that the average of the indexes will always be 100, that is, sum of the indexes should be 1200 for 12 months, and sum should be 400 for 4 quarterly data. If the sum of these 12 months percentages is not 1200, then the monthly percentage so obtained are adjusted by multiplying these by a suitable factor [1200 ÷ (sum of the 12 values)].

Example 12.15: The seasonal indexes of the sale of readymade garments in a store are given below:

Quarter	Seasonal Index
January to March	98
April to June	90
July to September	82
October to December	130

If the total sales of garments in the first quarter is worth Rs 1,00,000, determine how much worth of garments of this type should be kept in stock to meet the demand in each of the remaining quarters. [Delhi Univ., BCom, 1996]

Solution: Calculations of seasonal index for each quarter and estimated stock (in Rs) is shown in Table 12.11

Table 12.11 Calculation of Estimated Stock

<i>Quarter</i>	<i>Seasonal Index (SI)</i>	<i>Estimated Stock (Rs)</i>
Jan. —March	98	1,00,000.00
April —June	90	91,836.73*
July —Sept.	82	83,673.45
Oct. —Dec.	130	1,32,653.06

* These figures are calculated as follows:

$$\text{Seasonal index for second quarter} = \frac{\text{Figure for first quarter} \times \text{SI for second quarter}}{\text{SI for first quarter}}$$

$$\text{Seasonal index for third quarter} = \frac{\text{Figure for first quarter} \times \text{SI for third quarter}}{\text{SI for first quarter}}$$

Example 12.16: Use the method of monthly averages to determine the monthly indexes for the data of production of a commodity for the years 2002 to 2004.

<i>Month</i>	<i>2002</i>	<i>2003</i>	<i>2004</i>
January	15	23	25
February	16	22	25
March	18	28	35
April	18	27	36
May	23	31	36
June	23	28	30
July	20	22	30
August	28	28	34
September	29	32	38
October	33	37	47
November	33	34	41
December	38	44	53

Solution: Computation of seasonal index by average percentage method based on the data is shown in Table 12.12.

Table 12.12 Calculation of Seasonal Indexes

<i>Month</i>	<i>2002</i>	<i>2003</i>	<i>2004</i>	<i>Monthly Total for 3 Years</i>	<i>Monthly Averages for 3 Years</i>	<i>Percentage Average of Monthly Averages</i>
Jan.	15	23	25	63	21	70
Feb.	16	22	25	63	21	70
March	18	28	35	81	27	90
April	18	27	36	81	27	90
May	23	31	36	90	30	100
June	23	28	30	81	27	90
July	20	22	30	72	24	80
Aug.	28	28	34	90	30	100
Sept.	29	32	38	99	33	110
Oct.	33	37	47	117	39	130
Nov.	33	34	41	108	36	120
Dec.	38	44	53	135	45	150
				1080	360	1200

$$\text{Monthly Average} : 1080/20 = 90; \quad 360/12 = 30; \quad 1200/2 = 100$$

The average of monthly averages is obtained by dividing the total of monthly averages by 12. In column 7 each monthly average for 3 years have been expressed as a percentage of the averages. For example, the percentage for January is:

$$\text{Monthly index for January} = 21/30 = 70;$$

$$\text{February} = (21/30) \times 100 = 70$$

$$\text{March} = (27/30) \times 100 = 90, \text{ and so on}$$

Example 12.17: The data on prices (Rs in per kg) of a certain commodity during 2000 to 2004 are shown below:

Quarter	Years				
	2000	2001	2002	2003	2004
I	45	48	49	52	60
II	54	56	63	65	70
III	72	63	70	75	84
IV	60	56	65	72	66

Compute the seasonal indexes by the average percentage method and obtain the deseasonalized values.

Solution: Calculations for quarterly averages are shown in Table 12.13.

Table 12.13 Calculation Seasonal Indexes

Year	Quarters			
	I	II	III	IV
2000	45	54	72	60
2001	48	56	63	56
2002	49	63	70	65
2003	52	65	75	72
2004	60	70	84	66
Quarterly total	254	308	364	319
Quarterly average	50.8	61.6	72.8	63.8
Seasonal index	81.60	98.95	116.94	102.48

$$\text{Average of quarterly averages} = \frac{50.8 + 61.6 + 72.8 + 63.8}{4} = \frac{249}{4} = 62.25$$

$$\text{Thus, Seasonal index for quarter I} = \frac{50.8}{62.25} \times 100 = 81.60$$

$$\text{Seasonal index for quarter II} = \frac{61.6}{62.25} \times 100 = 98.95$$

$$\text{Seasonal index for quarter III} = \frac{72.8}{62.25} \times 100 = 116.94$$

$$\text{Seasonal index for quarter IV} = \frac{63.8}{62.25} \times 100 = 102.48$$

Deseasonalized Values Seasonal influences are removed from a time-series data by dividing the actual y value for each quarter by its corresponding seasonal index:

$$\text{Deseasonalized value} = \frac{\text{Actual quarterly value}}{\text{Seasonal index of corresponding quarter}} \times 100$$

The deseasonalized y values which are measured in the same unit as the actual values, reflect the collective influence of *trend*, *cyclical* and *irregular* forces. The deseasonalized values are given in Table 12.14.

Table 12.7 Calculation for Least Squares Equation

Year	Quarters			
	I	II	III	IV
2000	55.14	54.57	61.57	58.54
2001	58.82	56.59	53.87	54.64
2002	60.00	63.66	59.85	63.42
2003	63.72	65.68	64.13	70.25
2004	73.52	70.74	71.83	64.40

Limitations of the method of simple averages This method is the simplest of all the methods for measuring seasonal variation. However, the limitation of this method is that it assumes that there is no trend component in the series, that is, $C \cdot S \cdot I = 0$ or trend is assumed to have little impact on the time-series. This assumption is not always justified.

12.10.3 Ratio-to-Trend Method

This method is also known as the *percentage trend method*. This method is an improvement over the method of simple averages. Because here it is assumed that seasonal variation for a given month is a constant fraction of trend. The ratio-to-trend method isolates the seasonal factor when the following ratios are computed:

$$\frac{T \cdot S \cdot C \cdot I}{T} = S \cdot C \cdot I$$

The steps of the method are summarized as follows:

- (i) Compute the trend values by applying the least-squares method.
- (ii) Eliminate the trend value. In a multiplicative model the trend is eliminated by dividing the original data values by the corresponding trend values and multiplying these ratios by 100. The values so obtained are free from trend.
- (iii) Arrange the percentage data values obtained in Step (ii) according to months or quarters as the case may be for the various years.
- (iv) Find the monthly (or quarterly) averages of figures arranged in Step (iii) with any one of the usual measures of central tendency—arithmetic mean, median.
- (v) Find the grand average of monthly averages found in Step (iv). If the grand average is 100, then the monthly averages represent seasonal indexes. Otherwise, an adjustment is made by multiplying each index by a suitable factor [1200/(sum of the 12 values)] to get the final seasonal indexes.

Example 12.18: Quarterly sales data (Rs in million) in a super bazar are presented in the following table for a four-year period

Year	Quarters			
	I	II	III	IV
2000	60	80	72	68
2001	68	104	100	88
2002	80	116	108	96
2003	108	152	136	124
2004	160	184	172	164

Calculate the seasonal index for each of the four quarters using the ratio-to-trend method.

Solution: Calculations to obtain annual trend values from the given quarterly data using the method of least-squares are shown in Table 12.15.

Table 12.15 Calculation of Trend Values

Year	Yearly Total (1)	Yearly Average $y = (2)/4$	Deviation From Mid-Year x	x^2	xy	Trend Values \hat{y}
2000	280	70	-2	4	-140	64
2001	360	90	-1	1	-90	88
2002	400	100	0	0	0	0
2003	520	130	1	1	130	112
2004	680	170	2	4	340	160
		560		10	240	

Solving the following normal equations, we get

$$\Sigma y = na + b\Sigma x \quad 560 = 5a \quad \text{or } a = 112$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad 240 = 10b \text{ or } b = 24$$

Thus the yearly fitted trend line is: $y = 112 + 24x$. The value of $b = 24$ indicates yearly increase in sales. Thus the quarterly increment will be $24/4 = 6$.

To calculate quarterly trend values, consider first the year 2000. The trend value for this year is 64. This is the value for the middle of the year 2000, that is, half of the 2nd quarter and half of the 3rd quarter. Since quarterly increment is 6, the trend value for the 2nd quarter of 2000 would be $64 - (6/2) = 61$ and for the 3rd quarter it would be $64 + (6/2) = 67$. The value for the 1st quarter of 2000 would be $61 - 6 = 55$ and for the 4th quarter it would be $67 + 6 = 73$. Similarly, trend values of the various quarters of other years can be calculated as shown in Table 12.16.

Table 12.16 Quarterly Trend Values

Year	Quarters			
	I	II	III	IV
2000	55	61	67	73
2001	79	85	91	97
2002	103	109	115	121
2003	127	133	139	145
2004	151	157	163	169

After getting the trend values, the given data values in the time-series are expressed as percentages of the corresponding trend values in Table 12.16. Thus for the 1st quarter of 2000, this percentage would be $(60/55) \times 100 = 109.09$; for the 2nd quarter it would be $(80/61) \times 100 = 131.15$, and so on. Other values can be calculated in the same manner as shown in Table 12.17.

Table 12.17 Ratio-to-Trend Values

Year	Quarters			
	I	II	III	IV
2000	109.09	131.15	107.46	93.15
2001	86.08	122.35	109.89	90.72
2002	77.67	106.42	93.91	79.34
2003	85.04	114.29	97.84	85.52
2004	105.96	117.20	105.52	97.04
Total	463.84	591.41	514.62	445.77
Average	92.77	118.28	102.92	89.15
Adjusted seasonal index	92.02	117.33	102.09	88.43
			= 403.12	

The total of average of seasonal indexes is 403.12 (>400). Thus we apply the correction factor $(400/403.12) = 0.992$. Now each quarterly average is multiplied by 0.992 to get the adjusted seasonal index as shown in Table 12.17.

The seasonal index 92.02 in the first quarter means that on average sales trend to be depressed by the presence of seasonal forces to the extent of approx. $(100 - 92.02) = 7.98\%$. Alternatively, values of time series would be approx. $(7.98/92.02) \times 100 = 8.67\%$ higher had seasonal influences not been present.

12.10.4 Ratio-to-Moving Average Method

This method is also called the *percentage moving average method*. In this method, the original values in the time-series data are expressed as percentages of moving averages instead of percentages of trend values in the ratio-to-trend method. The steps of the method are summarized as follows:

- (i) Find the centred 12 monthly (or 4 quarterly) moving averages of the original data values in the time-series.
- (ii) Express each original data value of the time-series as a percentage of the corresponding centred moving average values obtained in Step (i). In other words, in a multiplicative time-series model, we get

$$\frac{\text{Original data values}}{\text{Trend values}} \times 100 = \frac{T \cdot C \cdot S \cdot I}{T \cdot C} \times 100 = (S \cdot I) \times 100\%$$

This implies that the ratio-to-moving average represents the seasonal and irregular components.

- (iii) Arrange these percentages according to months or quarter of given years. Find the averages over all months or quarters of the given years.
- (iv) If the sum of these indexes is not 1200 (or 400 for quarterly figures), multiply them by a correction factor $= 1200/(\text{sum of monthly indexes})$. Otherwise, the 12 monthly averages will be considered as seasonal indexes.

Example 12.19: Calculate the seasonal index by the ratio-to-moving method from the following data:

Year	Quarters			
	I	II	III	IV
2001	75	60	53	59
2002	86	65	63	80
2003	90	72	66	85
2004	100	78	72	93

Solution: Calculations for 4 quarterly moving averages and ratio-to-moving averages are shown in Table 12.18.

Table 12.18 Calculation of Ratio-to-Moving Averages

Year	Quarter	Original Values $Y = T.C.S.I$	4-Quarter Moving Total	4-Quarter Moving Average	$2 \times$ 4-Quarter Moving Average T.C	Ratio-to-Moving Average (Percent) $\frac{Y}{T.C} = (S.I)100\%$
2001	1	75	—	—	—	—
	2	60	248	507	63.375	54/63.375 = 85.20
	3	54	259	523	65.375	59/65.375 = 90.25
	4	59	264	537	67.125	128.12
2002	1	86	273	567	70.875	91.71
	2	65	294	592	74.000	85.13
	3	62	298	603	75.375	106.14
	4	80	305	613	76.625	117.43
2003	1	90	308	521	77.625	92.75
	2	72	313	636	79.500	83.02
	3	66	323	652	81.500	104.29
	4	85	329	664	84.750	92.03
2004	1	100	335	678	84.750	92.03
	2	78	343	—	—	—
	3	72	—	—	—	—
	4	93	—	—	—	—

Table 12.19 Calculation of Seasonal Index

Year	Quarters			
	I	II	III	IV..
2001	—	—	85.21	90.25
2002	128.12	91.71	85.13	106.14
2003	117.45	92.75	85.13	104.29
2004	120.48	92.03	—	—
Total	366.05	276.49	255.47	300.68
Seasonal average	91.51	69.13	63.87	75.17 = 299.66
Adjusted seasonal index	122.07	92.22	85.20	100.30 \equiv 400

The total of seasonal averages is 299.66. Therefore the corresponding correction factor would be $400/299.68 = 1.334$. Each seasonal average is multiplied by the correction factor 1.334 to get the adjusted seasonal indexes shown in Table 12.19.

Example 12.20: Calculate the seasonal indexes by the ratio-to-moving average method from the following data:

Year	Quarter	Actual Values $(Y = T.C.S.I)$	4-quarterly Moving Average	Year	Quarter	Given Values (Y)	4-quarterly Moving Average
2000	1	75	—	2002	1	90	76.625
	2	60	—		2	72	77.625
	3	54	63.375		3	66	79.500
	4	59	65.375		4	85	81.500
2001	1	86	67.125	2003	1	100	83.000
	2	65	70.875		2	78	84.750
	3	63	74.000		3	72	—
	4	80	75.375		4	93	—

Solution: Calculations of ratio-to-moving averages are shown in Table 12.20.

Table 12.20 Calculation of Seasonal Indexes

Year	Quarter	Actual	4-quarterly	Ratio to Moving
		Values (Y = T.C.S.I)	Moving (T.C)	Average (Percentage) $\frac{Y}{T.C} \times 100$
2000	1	75	—	—
	2	60	—	—
	3	54	63.375	85.21
	4	59	65.375	90.25
2001	1	86	67.125	128.12
	2	65	70.875	91.71
	3	63	74.000	85.14
	4	80	75.375	106.14
2002	1	90	76.625	117.46
	2	72	77.625	92.75
	3	66	79.500	83.02
	4	85	81.500	104.29
2003	1	100	83.000	120.84
	2	78	84.750	92.04
	3	72	—	—
	4	93	—	—

Rearranging the percentages to moving averages, the seasonal indexes are calculated as shown in Table 12.21.

Table 12.21 Seasonal Indexes

Year	Quarter (Percentages to Moving Averages)			
	1	2	3	4
2000	—	—	85.21	90.25
2001	128.12	91.71	85.14	106.14
2002	117.46	92.75	83.02	104.30
2003	120.48	92.04	—	—
Total	366.06	276.50	253.37	300.69
Average	122.02	92.17	84.46	100.23 = 398.88
Adjusted seasonal index	$\frac{122.02}{99.72} \times 100$ = 122.36	$\frac{92.17}{99.72} \times 100$ = 92.43	$\frac{84.46}{99.72} \times 100$ = 84.70	$\frac{100.23}{99.72} \times 100$ = 100.51 = 400

Since the total of average indexes is less than 400, the adjustment of the seasonal index has been done by calculating the grand mean value as follows:

$$\bar{\bar{x}} = \frac{122.02 + 92.17 + 84.46 + 100.23}{4} = 99.72$$

The seasonal average values are now converted into adjusted seasonal indexes using $\bar{\bar{x}} = 99.72$ as shown in Table 12.21.

Advantages and Disadvantages of Ratio-to-Moving Average Method This is the most widely used method for measuring seasonal variations because it eliminates both trend and cyclical variations from the time-series. However, if cyclical variations are not regular, then this method is not capable of eliminating them completely. Seasonal indexes calculated by this method will contain some effect of cyclical variations.

The only disadvantage of this method is that six data values at the beginning and the six data values at the end are not taken into consideration for calculation of seasonal indexes.

12.10.5 Link Relative Method

This method is also known as **Pearson's method**. The percentages obtained by this method are called **link relatives** as these link each month to the preceding one. The steps involved in this method are summarized below:

- Convert the monthly (or quarterly) data into link relatives by using the following formula:

$$\text{Link relative for a particular month} = \frac{\text{Data value of current month}}{\text{Data value of preceding month}} \times 100$$

- Calculate the average of link relatives of each month using either median or arithmetic mean.
- Convert the link relatives (L.R.) into chain relatives (C.R.) by using the formula:

$$\text{C.R. for a particular month} = \frac{[\text{L.R. of current month (or quarter)} \times \text{C.R. of preceding month (or quarter)}]}{100}$$

The C.R. for the first month (or quarter) is assumed to be 100.

- Compute the new chain relative for January (first month) on the basis of December (last month) using the formula:

$$\text{New C.R. for January} = \frac{\text{C.R. of January} \times \text{C.R. of December}}{100}$$

The new C.R. is usually not equal to 100 and therefore needs to be multiplied with the monthly correction factor

$$d = \frac{1}{12} (\text{New C.R. for January} - 100)$$

If the figures are given quarterly, then the correction factor would be

$$d = \frac{1}{4} (\text{New C.R. of first quarter} - 100)$$

The corrected C.R. for other months can be calculated by using the formula:

$$\text{Corrected C.R. for } k\text{th month} = \text{Original C.R. of } k\text{th month} - (k-1)d$$

where $k = 1, 2, 3, \dots, 12$

- Find the mean of the corrected chain index. If it is 100, then the corrected chain indexes represent the seasonal variation indexes. Otherwise divide the corrected C.R. of each month (or quarter) by the mean value of corrected C.R. and then multiply by 100 to get the seasonal variation indexes.

Example 12.21: Apply the method of link relatives to the following data and calculate seasonal indexes.

Year	Quarters			
	I	II	III	IV
1999	68	62	61	63
2000	65	58	56	61
2001	68	63	63	67
2002	70	59	56	62
2003	60	55	51	58

Solution: Computations of link relatives (L.R.) are shown in Table 12.22 by using the following formula:

$$\text{Link relative of any quarter} = \frac{\text{Data value of current quarter}}{\text{Data value of preceding quarter}} \times 100$$

Table 12.22 Computation of Link Relatives

Year	Quarters			
	I	II	III	IV
1999	—	91.18	98.39	103.28
2000	103.18	89.23	96.55	108.93
2001	111.48	92.65	100.00	106.35
2002	104.48	84.29	94.91	110.71
2003	96.78	91.67	92.73	113.73
Total of L.R.	415.92	449.02	482.58	543.00
Arithmetic mean of L.R.	103.98	89.80	96.52	108.60
Chain relatives (C.R.)	100	$\frac{89.80 \times 100}{100}$	$\frac{96.52 \times 89.80}{100}$	$\frac{108.60 \times 86.67}{100}$
		= 89.80	= 86.67	= 94.12

The new chain relatives for the first quarter on the basis of last quarter is calculated as follows:

$$\text{New C.R.} = \frac{\text{L.R. of first quarter} \times \text{C.R. of previous quarter}}{100} = \frac{103.98 \times 94.12}{100} = 97.9$$

Since new C.R. is not equal to 100, therefore we need to apply quarterly correction factor as:

$$\begin{aligned} d &= \frac{1}{4} (\text{New C.R. of first quarter} - 100) \\ &= \frac{1}{4} (97.9 - 100) = -0.53 \end{aligned}$$

Thus the corrected (or adjusted) C.R. for other quarters is shown in Table 12.23. For this we use the formula:

Corrected C.R. for kth quarter = Original C.R. of kth quarter - $(k - 1)d$
where $k = 1, 2, 3, 4$.

Table 12.23 Calculation of Link Relatives

Quarter	I	II	III	IV
Corrected C.R.	100	$89.80 - (-0.53)$ = 90.33	$86.67 - 2(-0.53)$ = 87.73	$94.13 - 3(-0.53)$ = 95.71
Seasonal indexes	$\frac{100}{93.44} \times 100$ = 107.02	$\frac{90.33}{93.44} \times 100$ = 96.67	$\frac{87.73}{93.44} \times 100$ = 93.89	$\frac{95.71}{93.44} \times 100$ = 102.42

$$\text{Mean of corrected C.R.} = \frac{100 + 90.33 + 87.73 + 95.71}{4} = 93.44$$

$$\text{Seasonal variation index} = \frac{\text{Corrected C.R.}}{\text{Mean of corrected C.R.}} \times 100$$

Example 12.22: Apply the method of link relatives to the following data and calculate the seasonal index:

Year	Quarters			
	I	II	III	IV
2000	45	54	72	60
2001	48	56	63	56
2002	49	63	70	65
2003	52	65	75	72
2004	60	70	84	86

Solution : Computations of link relatives (L.R.) using the following formula are shown in Table 12.24.

$$\text{L.R. of any quarter} = \frac{\text{Data value of current quarter}}{\text{Data value of preceding quarter}} \times 100$$

Table 12.24 Computation of Link Relatives

Year	Quarters			
	I	II	III	IV
2000	—	120	133.33	83.33
2001	80.00	116.67	112.50	88.89
2002	87.50	128.57	111.11	92.86
2003	80.00	125.00	115.38	96.00
2004	85.71	116.67	120.00	78.57
Total of L.R.	333.21	606.91	592.32	439.65
Arithmetic mean of L.R.	83.30	121.38	118.46	87.93
Chain relatives	100	$\frac{121.38 \times 100}{100}$	$\frac{118.46 \times 121.38}{100}$	$\frac{87.93 \times 143.78}{100}$
(C.R.)		= 121.38	= 143.78	= 126.42

The new chain relatives for the first quarter on the basis of the preceding quarter is calculated as follows:

$$\begin{aligned}\text{New C.R.} &= \frac{\text{L.R. of first quarter} \times \text{C.R. of previous quarter}}{100} \\ &= \frac{83.30 \times 126.42}{100} = 105.30\end{aligned}$$

Since the new C.R. is more than 100, therefore we need to apply a quarterly correction factor as :

$$\begin{aligned}d &= \frac{1}{4} (\text{New C.R. of first quarter} - 100) \\ &= \frac{1}{4} (105.30 - 100) = 1.325\end{aligned}$$

Thus the corrected (or adjusted) C.R. for other quarters is shown in Table 12.25. For this we use the formula

Corrected C.R. for k th quarter = Original C.R. of k th quarter $- (k - 1)d$
where $k = 1, 2, 3, 4$.

Table 12.25 Corrected C.R.

Quarters	I	II	III	IV
Corrected C.R.	100	$121.38 - 1.32$ = 120.06	$143.78 - 2(1.32)$ = 141.14	$126.42 - 3(1.32)$ = 122.46
Seasonal indexes	$\frac{100}{120.92} \times 100$ = 82.70	$\frac{120.06}{120.92} \times 100$ = 99.30	$\frac{141.14}{120.92} \times 100$ = 116.72	$\frac{122.46}{120.92} \times 100$ = 101.27

$$\text{Mean of corrected C.R.} = \frac{100 + 120.06 + 141.14 + 122.46}{4} = 120.92$$

$$\text{Seasonal variation index} = \frac{\text{Corrected C.R.}}{\text{Mean of corrected C.R.}} \times 100$$

Advantages and Disadvantages of Link Relative Method This method is much simpler than the ratio-to-trend or the ratio-to-moving average methods. In this method the L.R. of the

first quarter (or month) is not taken into consideration as compared to ratio-to-trend method, where 6 values each at the beginning and at the end periods (month) are lost.

This method eliminates the trend but it is possible only if there is a straight line (linear) trend in the time-series—which is generally not formed in business and economic series.

12.11 MEASUREMENT OF CYCLICAL VARIATIONS—RESIDUAL METHOD

As mentioned earlier that a typical time-series has four components: secular trend (T), seasonal variation (S), cyclical variation (C), and irregular variation (I). In a multiplicative time-series model, these components are written as:

$$y = T \cdot C \cdot S \cdot I$$

The deseasonalization data can be adjusted for trend analysis by dividing these by the corresponding trend and seasonal variation values. Thus we are left with only cyclical (C) and irregular (I) variations in the data set as shown below:

$$\frac{y}{T \cdot S} = \frac{T \cdot C \cdot S \cdot I}{T \cdot S} = C \cdot I$$

The moving averages of an appropriate period may be used to eliminate or reduce the effect of irregular variations and thus left behind only the cyclical variations.

The procedure of identifying cyclical variation is known as the *residual method*. Recall that cyclical variations in time-series tend to oscillate above and below the secular trend line for periods longer than one year. The steps of residual method are summarized as follows:

- (i) Obtain seasonal indexes and deseasonalized data.
- (ii) Obtain trend values and express seasonalized data as percentages of the trend values.
- (iii) Divide the original data (y) by the corresponding trend values (T) in the time-series to get S. C. I. Further divide S. C. I by S to get C. I.
- (iv) Smooth out irregular variations by using moving averages of an appropriate period but of short duration, leaving only the cyclical variation.

12.12 MEASUREMENT OF IRREGULAR VARIATIONS

Since irregular variations are random in nature, no particular procedure can be followed to isolate and identify these variations. However, the residual method can be extended one step further by dividing C. I by the cyclical component (C) to identify the irregular component (I).

Alternately, trend (T), seasonal (S), and cyclical (C) components of the given time-series are estimated and then the residual is taken as the irregular variation. Thus, in the case of multiplicative time-series model, we have

$$\frac{Y}{T \cdot C \cdot S} = \frac{T \cdot C \cdot S \cdot I}{T \cdot C \cdot S} = I$$

where S and C are in fractional form and not in percentages.

Conceptual Questions 12B

- 15. (a) Under what circumstances can a trend equation be used to forecast a value in a series in the future? Explain.
- (b) What are the advantages and disadvantages of trend analysis? When would you use this method of forecasting?
- 16. What effect does seasonal variability have on a time-series? What is the basis for this variability for an economic time-series?
- 17. What is measured by a moving average? Why are 4-quarter and 12-month moving averages used to develop a seasonal index?

- 18.** Briefly describe the moving average and least squares methods of measuring trend in time-series.
 [CA, May 1997]
- 19.** Explain the simple average method of calculating indexes in the context of time-series analysis.
- 20.** Distinguish between ratio-to-trend and ratio-to-moving average as methods of measuring seasonal variations. Which is better and why?
- 21.** Distinguish between trend, seasonal variations, and cyclical variations in a time-series. How can trend be isolated from variations?
- 22.** Describe any two important methods of trend measurement, and examine critically the merits and demerits of these methods.
- 23.** Why do we deseasonalize data? Explain the ratio-to-moving average method to compute the seasonal index.
- 24.** Explain the following:

- (a) ‘... , the business analyst who uses moving averages to smoothen data, while in the process of trying to discover business cycles, is likely to come up with some non-existent cycles’.
- (b) ‘Despite great limitations of statistical forecasting, the forecasting techniques are invaluable to the economist, the businessman, and the Government.’
- 25.** ‘A 12-month moving average of time-series data removes trend and cycle’. Do you agree ? Why or why not?
- 26.** Why do we deseasonalize data? Explain the ratio-to-moving average method to compute the seasonal index.
- 27.** Explain the methods of fitting of the quadratic and exponential curves. How would you use the fitted curves for forecasting?
- 28.** ‘A key assumption in the classical method of time-series analysis is that each of the component movements in the time-series can be isolated individually from a series’. Do you agree with this statement? Does this assumption create any limitation to such analysis?

Self-Practice Problems 12C

- 12.19** Apply the method of link relatives to the following data and calculate seasonal indexes.

Quarter	1999	2000	2001	2002	2003
I	6.0	5.4	6.8	7.2	6.6
II	6.5	7.9	6.5	5.8	7.3
III	7.8	8.4	9.3	7.5	8.0
IV	8.7	7.3	6.4	8.5	7.1

- 12.20** A company estimates its sales for a particular year to be Rs 24,00,000. The seasonal indexes for sales are as follows:

Month	Seasonal Index	Month	Seasonal Index
January	75	July	102
February	80	August	104
March	98	September	100
April	128	October	102
May	137	November	82
June	119	December	73

Using this information, calculate estimates of monthly sales of the company. (Assume that there is no trend).

[Osmania Univ., MBA, 1997]

- 12.21** Calculate the seasonal index from the following data using the average method:

Year	Quarter			
	I	II	III	IV
2000	72	68	80	70
2001	76	70	82	74
2002	74	66	84	80
2003	76	74	84	78
2004	78	74	86	82

[Kerala Univ., BCom, 1996]

- 12.22** Calculate seasonal index numbers from the following data:

Year	Quarter			
	I	II	III	IV
1998	108	130	107	93
1999	86	120	110	91
2000	92	118	104	88
2001	78	100	94	78
2002	82	110	98	86
2003	106	118	105	98

- 12.23** Calculate seasonal index for the following data by using the average method:

Year	Quarters			
	I	II	III	IV
2000	72	68	80	70
2001	76	70	82	74
2002	74	66	84	80
2003	76	74	84	78
2004	78	74	86	82

- 12.24** On the basis of quarterly sales (Rs in lakh) of a certain commodity for the years 2003—2004, the following calculations were made:

Trend : $y = 20 + 0.5t$ with origin at first quarter of 2003

where t = time unit (one quarter),

y = quarterly sales (Rs in lakh)

Seasonal variations:

Quarter : 1 2 3 4

Seasonal index : 80 90 120 110

Estimate the quarterly sale for the year 2003 using multiplicative model.

Hints and Answers

12.19

Year	Quarters			
	I	II	III	IV
1999	—	108.3	120.0	111.5
2000	62.1	146.3	106.3	89.9
2001	93.2	95.6	143.1	68.8
2002	112.5	80.6	129.3	113.3
2003	77.6	110.6	109.6	88.8
Arithmetic average	$\frac{345.4}{4} = 86.35$	$\frac{541.4}{5} = 108.28$	$\frac{608.3}{5} = 121.66$	$\frac{469.3}{5} = 93.86$
Chain relatives	100	$\frac{100 \times 108.28}{100} = 108.28$	$\frac{121.66 \times 108.28}{100} = 131.73$	$\frac{93.86 \times 131.73}{100} = 123.65$
Corrected chain relatives	100	$108 - 1.675 = 106.325$	$131.73 - 3.35 = 128.38$	$123.64 - 5.025 = 118.615$
Seasonal indexes	$\frac{100 \times 100}{113.4} = 88.18$	$\frac{106.605}{113.4} \times 100 = 94.01$	$\frac{128.38}{113.4} \times 100 = 113.21$	$\frac{118.615}{113.4} \times 100 = 104.60$

- 12.20** Seasonal indexes are usually expressed as percentages. The total of all the seasonal indexes is 1200.

$$\text{Seasonal effect} = \text{Seasonal index} + 100$$

The yearly sales being Rs 24,00,000, the estimated monthly sales for a specified month:

$$\begin{aligned}\text{Estimated sales} &= \frac{\text{Annual sales}}{12} \times \text{Seasonal effect} \\ &= \frac{24,00,000}{12} \times \text{Seasonal effect} \\ &= 2,00,000 \times \text{Seasonal effect}\end{aligned}$$

(1)	(2)	Month	Seasonal Index	Seasonal Effect (3) = (2) ÷ 100	Estimated Sales (4) = (3) × 2,00,000
January	75		0.75		1,50,000
February	80		0.80		1,60,000
March	98		0.98		1,96,000
April	128		1.28		2,56,000
May	137		1.37		2,74,000
June	119		1.19		2,38,000
July	102		1.02		2,04,000
August	104		1.04		2,08,000
September	100		1.00		2,00,000
October	102		1.02		2,04,000
November	82		0.82		1,64,000
December	73		0.73		1,46,000
		1200	12.00		24,00,000

12.21

Year	Quarters			
	I	II	III	IV
2000	72	68	80	70
2001	76	70	82	74
2002	74	66	84	80
2003	76	74	84	78
2004	78	74	86	82
Total	376	352	416	384
Average	75.2	70.4	83.2	76.8
Seasonal index	98.43	92.15	108.9	100.52

$$\begin{aligned}\text{Grand average} &= \frac{75.2 + 70.4 + 83.2 + 76.8}{4} \\ &= \frac{305.6}{4} = 76.4\end{aligned}$$

Seasonal index for quarter

$$k = \frac{\text{Average of quarter } k}{\text{Grand average}} \times 100$$

12.22

Year	Quarters			
	I	II	III	IV
1998	108	130	107	93
1999	86	120	110	91
2000	92	118	104	88
2001	78	100	94	78
2002	82	110	98	86
2003	106	118	105	98
Total	552	696	618	534
Average	92	116	103	89
Seasonal	$\frac{92}{100} \times 100$	$\frac{116}{100} \times 100$	$\frac{103}{100} \times 100$	$\frac{89}{100} \times 100$
Index	= 92	= 116	= 103	= 89

Sales in different quarters:

I: Rs 20,000; II: $20,000 \times 1.16 = \text{Rs } 23,200$;

III: $20,000 \times 1.03 = \text{Rs } 20,600$;

IV: $20,000 \times 0.89 = \text{Rs } 17,800$

12.23

Year	Quarters			
	I	II	III	IV
2000	72	68	80	70
2001	76	70	82	74
2002	74	66	84	80
2003	76	74	84	78
2004	78	74	86	82
Total	376	352	416	384
Average	75.2	70.4	83.2	76.8
Seasonal	$\frac{75.2}{76.4} \times 100$	$\frac{70.4}{76.4} \times 100$	$\frac{83.2}{76.4} \times 100$	$\frac{76.8}{76.4} \times 100$
Index	= 98.43	= 92.15	= 108.90	= 100.52

12.24

Quarter of 2003	Time Unit	Trend (T) Values	Seasonal Effect or Seasonal Index	Estimated Sales (Rs in lakh)	
				(S)	T · S
1	4	$20 + 0.5 \times 4 = 22.0$	0.80	17.60	
2	5	$20 + 0.5 \times 5 = 22.5$	0.90	20.25	
3	6	$20 + 0.5 \times 6 = 23.0$	1.20	27.60	
4	7	$20 + 0.5 \times 7 = 23.5$	1.10	25.85	

Formulae Used

1. Secular trend line

- Linear trend model

$$y = a + bx$$

$$\bullet \text{ where } a = \bar{y} - b \bar{x}; \quad b = \frac{\sum xy - n \bar{x} \bar{y}}{\sum x^2 - n (\bar{x})^2}$$

- Exponential trend model

$$y = ab^x;$$

$$\log a = \frac{1}{n} \sum \log y; \quad \log b = \frac{\sum x \log y}{\sum x^2}$$

- Parabolic trend model

$$y = a + bx + cx^2$$

- where $a = \frac{\sum y - c \sum x^2}{n}; \quad b = \frac{\sum xy}{\sum x^2}$

$$c = \frac{n \sum x^2 y - \sum x^2 \sum y}{n \sum x^4 - (\sum x^2)^2}$$

2. Moving average

$$MA_{t+1} = \frac{\sum \{D_t + D_{t-1} + \dots + D_{t-n+1}\}}{n}$$

where $t = \text{current time period}$

D = actual data value

n = length of time period

3. Simple exponential smoothing

$$\begin{aligned} F_t &= F_{t-1} + \alpha(D_{t-1} - F_{t-1}) \\ \text{where } F_t &= \text{current period forecast} \\ F_{t-1} &= \text{previous period forecast} \\ \alpha &= \text{a weight (}0 \leq \alpha \leq 1\text{)} \\ D_{t-1} &= \text{previous period actual demand} \end{aligned}$$

4. Adjusted exponential smoothing

$$\begin{aligned} (F_t)_{\text{adj}} &= F_t + \frac{1-\beta}{\beta} T_t \\ \text{where } \beta &= \text{smoothing constant for trend} \\ T_t &= \text{exponential smoothed trend factor} \end{aligned}$$

Review Self-Practice Problems

- 12.25** A sugar mill is committed to accepting beets from local producers and has experienced the following supply pattern (in thousands of tons/year and rounded).

Year	Tonnes	Year	Tonnes
1990	100	1995	400
1991	100	1996	400
1992	200	1997	600
1993	600	1998	800
1994	500	1999	800

The operations manager would like to project a trend to determine what facility additions will be required by 2004.

- (a) Sketch a freehand curve and extend it to 2004. What would be your 2004 forecast based upon the curve?
 - (b) Compute a three-year moving average and plot it as a dotted line on your graph.
- 12.26** Use the data of Problem 12.25 and the normal equations to develop a least squares line of best fit. Omit the year 1990.
- (a) State the equation when the origin is 1995.
 - (b) Use your equation to estimate the trend value for 2004.

- 12.27** A forecasting equation is of the form:

$$\hat{y}_c = 720 + 144x$$

[2003 = 0, x unit = 1 year, y = annual sales]

- (a) Forecast the annual sales rate for 2003 and also for one year later.
- (b) Change the time (x) scale to months and forecast the annual sales rate at July 1, 2003, and also at one year later.
- (c) Change the sales (y) scale to monthly and forecast the monthly sales rate at July 1, 2003, and also at one year later.

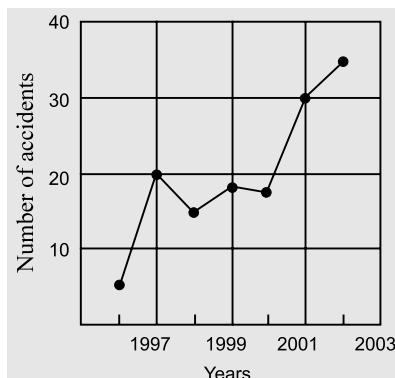
- 12.28** Data collected on the monthly demand for an item were as shown below:

January	100
February	90
March	80
April	150
May	240
June	320
July	300
August	280
September	220

- (a) What conclusion can you draw with respect to the length of moving average versus smoothing effect?
- (b) Assume that the 12-month moving average centred on July was 231. What is the value of the ratio-to-moving average that would be used in computing a seasonal index?

- 12.29** The data shown in the table below gives the number of lost-time accidents over the past seven years in a cement factory:

Year	Number Employees (in '000)	Number Accidents
1996	15	5
1997	12	20
1998	20	15
1999	26	18
2000	35	17
2001	30	30
2002	37	35



- (a) Use the normal equations to develop a linear time-series equation forecasting the number of accidents.
 (b) Use your equation to forecast the number of accidents in 2005.

12.30 Consider the following time-series data:

Week :	1	2	3	4	5	6
Value :	8	13	15	17	16	9

- (a) Develop a 3-week moving average for this time-series. What is the forecast for week 7?
 (b) Use $\alpha = 0.2$ to compute the exponential smoothing values for the time-series. What is the forecast for week 7?

12.31 Admission application forms data (1000's) received by a management institute over the past 6 years are shown below:

Year :	1	2	3	4	5	6
Application forms :	20.5	20.2	19.5	19.0	19.1	18.8

Develop the equation for the linear trend component of this time-series. Comment on what is happening to admission forms for this institution.

12.32 Consider the following time-series data:

Quarter	Year		
	1	2	3
1	4	6	7
2	2	3	6
3	3	5	6
4	5	7	8

- (a) Show the 4-quarter moving average values for this time-series.
 (b) Compute seasonal indexes for the 4 quarters.

12.33 Below are given the figures of production (in million tonnes) of a cement factory:

Hints and Answers

12.25 (a) Forecasts is around 1200 (thousand) tonnes
 (b) Averages are: 133, 300, 433, 500, 433, 466, 600 and 733.

12.26 (a) $\hat{y} = 489 + 75x$ [1995 = 0, x = years, y = tonnes in thousand]
 (b) 11,64,000 tonnes

12.27 (a) 720 units when $x = 0$, 864 units when $x = 1$.

- (b) $\hat{y} = 720 + 12x$ [July 1, 2003 = 0; x unit = 1 month; y = annual sales rates in units]
 720 units per year; 864 units per year.
 (c) $\hat{y} = 60 + x$ [July 1, 2003 = 0, x unit = 1 month; y = monthly sales rates in units]
 60 units per month; 72 units per month.

Year : 1990 1992 1993 1994 1995 1996 1999

Production : 77 88 94 85 91 98 90

- (a) Fit a straight line trend by the 'least squares method' and tabulate the trend values.
 (b) Eliminate the trend. What components of the time series are thus left over?
 (c) What is the monthly increase in the production of cement? [Sukhadia Univ., MBA, 1999]

12.34 The sale of commodity in tonnes varied from January 2000 to December, 2000 in the following manner:

280	300	280	280	270	240
230	230	220	200	210	200

Fit a trend line by the method of semi-averages.

12.35 Fit a parabolic curve of the second degree to the data given below and estimate the value for 2002 and comment on it.

Year : 1996 1997 1998 1999 2000
 Sales

(Rs in '000) : 10 12 13 10 8

12.36 Given below are the figures of production of a sugar (in 1000 quintals) factory:

Year : 1991 1992 1993 1994 1995 1996 1997

Production : 40 45 46 42 47 49 46

Fit a straight line trend by the method of least squares and estimate the value for 2001.

[MBA, MD Univ., 1998]

12.37 The following table gives the profits (Rs in thousand) of a concern for 5 years ending 1996.

Year : 1996 1997 1998 1999 2000

Profits : 1.6 4.5 13.8 40.2 125.0

Fit an equation of the type $y = ab^x$.

12.28 (a) Longer average yield more smoothing; (b) 1.3

12.29 (a) $\hat{y} = 20 + 4x$ [$1999 = 0$, x = years; y = number of accidents]; (b) 44

12.30 (a)

Week (1)	Values (2)	Forecast (3)	Forecast Error (4) = (2) - (3)	Squared Forecast Error
1	8	—	—	—
2	13	—	—	—
3	15	—	—	—
4	17	12	5	25
5	16	15	1	1
6	9	16	-7	49

Forecast for week 7 is: $(17 + 16 + 9)/3 = 14$.

(b)

Week (t)	Values y_t	Forecast F_t	Forecast Error $y_t - F_t$	Squared Error $(y_t - F_t)^2$
1	8	—	—	—
2	13	8.00	5.00	25.00
3	15	9.00	6.00	36.00
4	17	10.20	6.80	46.24
5	16	11.56	4.44	19.71
6	9	12.45	-3.45	11.90
				138.85

Forecast for week 7 is: $0.2(9) + (1 - 0.2)(12.45) = 11.76$.

$$12.31 \Sigma x = 21, \Sigma x^2 = 91, \Sigma y = 117.1, \Sigma xy = 403.7$$

$$b = \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2} = \frac{6(403.7) - 21 \times 117.2}{6 \times 91 - (21)^2} = -0.3714$$

$$a = \bar{y} - b\bar{x} = 19.5167 - (-0.3514)(3.5) = 20.7466$$

$$\hat{y} = 20.7466 - 0.3514x.$$

Enrolment appears to be decreasing by about 351 students per year.

12.32 (a)

Year	Quarter	Value y	4-quarter Moving Average	Centred Moving Average
1	1	4	3.50	3.750
	2	2		
	3	3		
	4	5		
2	1	6	4.25	4.125
	2	3		
	3	5		
	4	7		
3	1	7	6.25	6.375
	2	6		
	3	6		
	4	8		

(b)

Year	Quarter	Value y	Centered Moving Average	Seasonal Component
1	1	4	3.750	0.8000
	2	2		
	3	3		
	4	5		
2	1	6	4.500	1.3333
	2	3		
	3	5		
	4	7		
3	1	7	6.25	1.0000
	2	6		
	3	6		
	4	8		

Quarter	Seasonal-Irregular Component Values	Seasonal Index
1	1.333, 1.0980	1.2157
2	0.6000, 0.9057	0.7529
3	0.8000, 0.9302	0.8651
4	1.2121, 1.1915	1.2018
		4.0355

$$\text{Adjusted for seasonal index} = \frac{4}{4.0355} = 0.9912.$$

12.33 (a)

Year	Time Period	Production (in m. tonnes)		Deviation From 1994		Trend Values
		y	x	xy	x	\hat{y}
1990	-4	77	-4	-308	16	83.299
1992	-2	88	-2	-176	4	86.051
1993	-1	94	-1	-94	1	87.427
1994	0	85	0	0	0	88.803
1995	1	91	1	91	1	90.179
1996	2	98	2	196	4	91.555
1999	5	90	5	450	25	95.683
		623	1	159	51	

Solving the normal equations

$$\Sigma y = na + b\Sigma x \quad 623 = 7a + b$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad 159 = a + 5b$$

we get $a = 88.803$ and $b = 1.376x$. Thus

$$\hat{y} = a + bx = 88.803 + 1.376x$$

Substituting $x = -4, -2, -1, 0, 1, 2, 5$ to get trend values as shown above in the table.

(b) After eliminating the trend, we are left with S, C, and I components of time-series.

(c) Monthly increase in the production of cement is given by $b/12 = 1.376/12 = 0.115$.

12.34

Month (in tonnes)	Sales
January	280
February	300
March	280
April	280
May	270
June	240
July	230
August	230
September	220
October	200

Plot 275 and 215 in the middle of March-April 2000 and that of September-October 2000. By joining these two points we get a trend line which describes the given data.

12.35

Year	Sales	Period				
		y	x	xy	x^2	x^2y
1996	10	-2	-20	4	40	16
1997	12	-1	-12	1	12	1
1998	13	0	0	0	0	0
1999	10	1	10	1	10	1
2000	8	2	16	4	32	16
	53	0	-6	10	94	34

Parabolic trend line : $y = a + bx + bx^2$

$$a = \frac{\Sigma y - c\Sigma x^2}{n} = \frac{53 - 0.857 \times 10}{5} = 8.886$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{-6}{10} = -0.6 ;$$

$$c = \frac{n\Sigma x^2y - \Sigma x^2 \Sigma y}{n\Sigma x^4 - (\Sigma x^2)^2} = \frac{5(94) - 10(53)}{5(34) - (10)^2} = -0.857$$

$$\therefore y = 8.886 - 0.6x - 0.857x^2$$

$$\text{For } 2002, x = 4; \quad y = 8.886 - 0.6(4) - 0.857(4)^2 \\ = -7.226$$

12.36

Year	Production ('000 qts)	Deviations from 1994			
		y	x	xy	x^2
1991	40	-3	-120	9	
1992	45	-2	-90	4	
1993	46	-1	-46	1	
1994	42	0	0	0	
1995	47	1	47	1	
1996	49	2	98	4	
1997	46	3	138	9	
	315	0	27	28	

$$\hat{y} = a + bx; \quad a = \Sigma y/n = 315/7 = 45;$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{27}{28} = 0.964$$

$$\hat{y} = 45 + 0.964x$$

$$y_{2001} = 45 + 0.964(7) = 45 + 6.748 = 51.748$$

12.37

Year	Profits	x y	Log y	x^2	$x . Log y$
1996	1.6	-2	0.2041	4	-0.4082
1997	4.5	-1	0.6532	1	-0.6532
1998	13.8	0	1.1399	0	0
1999	40.2	1	1.6042	1	1.6042
2000	125.0	2	2.0969	4	4.1938
	185.1	0	5.6983	10	4.7366

Trend line: $y = ab^x$ or $\log y = \log a + x \log b$

$$\text{where } \log a = \frac{\Sigma \log y}{n} = \frac{5.6983}{5} = 1.1397;$$

$$\log b = \frac{\Sigma x \log y}{\Sigma x^2} = \frac{4.7366}{10} = 0.474$$

Thus $\log y = 1.1397 + 0.474x$.

*And in such indexes ..., there
is seen the baby figure of the
gaint mass of things to
come.*

—William Shakespeare

Index Numbers

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- explain the purpose of index numbers.
- compute indexes to measure price changes and quantity changes over time.
- revise the base period of a series of index numbers
- explain and derive link relatives
- discuss the limitations of index number construction

13.1 INTRODUCTION

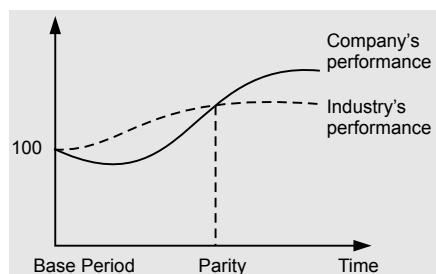
We know that most values change and therefore may want to know-how much change has taken place over a period of time. For example, we may want to know-how much the prices of different items essential to a household have increased or decreased so that necessary adjustments can be made in the monthly budget. An organization may be concerned with the way in which prices paid for raw materials, annual income and profit, commodity prices, share prices, production volume, advertising budget, wage bills, and so on, have changed over a period of time. However, while prices of a few items may have increased, others may have decreased over a given period of time. Consequently in all such situations, an average measure needs to be defined to compare such differences from one time period to another. *Index numbers* are yardsticks for describing such difference.

An *index number* can be defined as a relative measure describing the average changes in any quantity over time. In other words, an index number measures the changing value of prices, quantities, or values over a period of time in relation to its value at some fixed point in time, called the *base period*. This resulting ratio of the current value to a base value is multiplied by 100 to express the index as a percentage. Since an index number is constructed as a ratio of a measure taken during one time period to that same measure taken during another time period (called base period), it has no unit and is always expressed as a percentage term as follows:

$$\text{Index number} = \frac{\text{Current period value}}{\text{Base period value}} \times 100$$

Indexes may be based at any convenient period, which is occasionally adjusted, and these are published at any convenient frequency. Examples of some indexes are:

Figure 13.1
Graph of Two Indexes



Daily	Stock market prices
Monthly	Unemployment figures
Yearly	Gross National Product (GNP)

Index numbers were originally developed by economists for monitoring and comparing different groups of goods. For decision-making in business, it is sometimes essential to understand and manipulate different published index series and to construct one's own index series. This index series can be compared with a national one and/or with competitor's. For example, a cement company could construct an index of its own sales and production volumes and compare it to the index of the cement industry. A graph of two indexes will provide, at a glace, a view of a company's performance within the industry, as shown in Fig. 13.1.

13.2 INDEX NUMBER DEFINED

Definition of index numbers can be classified into the following three broad categories:

1. A measure of change

- It is a numerical value characterizing the change in complex economic phenomena over a period of time or space. —Maslow
- An index number is a quantity which, by reference to a base period, shows by its variations the changes in the magnitude over a period of time. In general, index numbers are used to measure changes over time in magnitudes which are not capable of direct measurement. —John I. Raffin
- An index number is a statistical measure designed to show changes in variables or a group of related variables with respect to time, geographic location or other characteristics. —Speigel
- Index number is a single ratio (usually in percentages) which measures the combined (i.e., averaged) change of several variables between two different times, places or situations. —A. M. Tuttle

2. A device to measure change

- Index numbers are devices measuring differences in the magnitude of a group of related variables. —Corxton and Cowden
- An index number is a device which shows by its variation the changes in a magnitude which is not capable of accurate measurement in itself or of direct valuation in practice. —Wheldom

3. A series representing the process of change

- Index numbers are series of numbers by which changes in the magnitude of a phenomenon are measured from time to time or place to place. —Horace Secris
- A series of index numbers reflects in its trend and fluctuations the movements of some quantity of which it is related. —B. L. Bowley
- An index number is a statistical measure of fluctuations in a variable arranged in the form of a series, and using a base period for making comparisons. —L. J. Kaplan

13.3 TYPES OF INDEX NUMBERS

Index numbers are broadly classified into three categories: (i) price indexes, (ii) quantity indexes, and (iii) value indexes. A brief description of each of these is as follows:

Price Indexes These indexes are of two categories:

- Single price index
- Composite prices index

The single price index measures the percentage change in the current price per unit of a product to its base period price. To facilitate comparisons with other years, the actual per unit price is converted into a *price relative*, which expresses the unit price in each period as a percentage of unit price in a base period. Price relatives are very helpful to understand and interpret changing economic and business conditions over time. Table 13.1 illustrates the calculations of price relatives,

Table 13.1: Calculation of Price Index (Base year = 1996)

Year (I)	Total Wage Bill (Rs millions) (2)	Ratio (3) = (2)/11.76	Price Index or Percentage Relative (4) = (3) × 100
2000	11.76	11.76/11.76 = 1.0	100.0
2001	12.23	12.23/11.76 = 1.039	103.9
2002	12.84	12.84/11.76 = 1.091	109.1
2003	13.35	13.35/11.76 = 1.135	113.5
2004	13.82	13.82/11.76 = 1.175	117.5

From Table 13.1, it is observed that the price relative of 113.5 in 2003 shows a increase of 13.5% in wage bill compared to the base year 2000.

A composite price index measures the average price change for a basket of related items from a base period to the current period. For example, the *wholesale price index* reflects the general price level for a group of items (or a basket of items) taken as a whole.

The *retail price index* reflects the general changes in the retail prices of various items including food, housing, clothing, and so on. In India, the Bureau of Labour statistics, publishes retail price index. The consumer price index, a special type of retail price index, is the primary measure of the cost of living in a country. The consumer price index is a weighted average price index with fixed weights. The weightage applied to each item in the basket of items is derived from the urban and rural families.

Quantity Index A quantity index measures the relative changes in quantity levels of a group (or basket) of items consumed or produced, such as agricultural and industrial production, imports and exports, between two time periods. The method of constructing quantity indexes is the same as that of price index except that the quantities are vary from period to period.

Quantity index: An index that is constructed to measure changes in quantities over time.

The two most common quantity indexes are the weighted *relative of aggregates* and the weighted average of quantity relative index.

Value Index A value index measures the relative changes in total monetary worth of an item, such as inventories, sales, or foreign trade, between the current and base periods. The value of an item is determined by multiplying its unit price by the quantity under consideration. The value index can also be used to measure differences in a given variable in different locations. For example, the comparative cost of living shows that in terms of cost of goods and services, it is cheaper to live in a small city than in metro cities.

Special Purpose Indexes A few index numbers such as industrial production, agricultural production, productivity, etc. can also be constructed separately depending on the nature and degree of relationship between groups and items.

- Index number, almost alone in the domain of social sciences, may truly be called an exact science, if it be permissible to designate as science the theoretical foundations of a useful art.

—Irving Fisher.

13.4 CHARACTERISTICS AND USES OF INDEX NUMBERS

Based on the definitions and types of index numbers discussed earlier in this chapter, the following characteristics and uses of index number emerge.

13.4.1 Characteristics of Index Numbers

Consumer price index: A price index that uses the price changes in a market basket of consumer goods and services to measure the changes in consumer prices over time.

1. **Index numbers are specialized averages:** According to R. L. Corner, '*An index number represents a special case of an average, generally weighted average, compiled from a sample of items judged to be representative of the whole*'.

'Average' is a single figure representing the characteristic of a data set. This figure can be used as a basis for comparing two or more data sets provided the unit of measurement of observations in all sets is the same. However, index numbers which are considered as a special case of average can be used for comparison of two or more data sets expressed in different units of measurement.

The consumer price index, for example, which represents a price comparison for a group of items—food, clothing, fuel, house rent, and so on, are expressed in different units. An average of prices of all these items expressed in different units is obtained by using the technique of price index number calculation.

2. **Index numbers measure the change in the level of phenomena in percentages:** Since index numbers are considered as a special case of an average, these are used to represent, in one single figure, the increase or decrease (expressed in terms of percentage) in the value of a variable. For example, a quantity index number of 110 for cars sold in a given year when compared with that of a base year would mean that cars sales in the given year were 10 per cent higher than in the base year (value of index number in base period is always equal to 100). Similarly, a quantity index number of 90 in a given year would indicate that the number of cars sold in the given were 10 per cent less than in the base year.
3. **Index numbers measure changes in a variety of phenomena which cannot be measured directly:** According to Bowley, '*Index numbers are used to measure the changes in some quantity which we cannot observe directly . . .*'

It is not possible, for example, to directly measure the changes in the import-export activities of a country. However, it is possible to study relative changes in import and export activities by studying the variations in factors such as raw materials available, technology, competitors, quality, and other parameters which affect import and export, and are capable of direct measurement. Similarly, cost of living cannot be measured in quantitative terms directly, we can only study relative changes in it by studying the variations in certain other factors connected to it.

4. **Index numbers measure the effect of changes in relation to time or place:** Index numbers are used to compare changes which take place over periods of time, between locations, and in categories. For example, cost of living may be different at two different places at the same or cost of living in one city can be compared across two periods of time.

13.4.2 Uses of Index Numbers

According to G. Simpson and F. Kafka '*Today Index numbers are one of the most widely used statistical tools. They are used to feel the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies*'. Other important uses of index number can be summarized as follows:

1. **Index numbers act as economic barometers:** A barometer is an instrument that is used to measure atmospheric pressure. Index numbers are used to feel the pressure of the economic and business behaviour, as well as to measure ups and downs in the general economic condition of a country. For example, the composite index number of indexes of prices, industrial output, foreign exchange reserves, and bank deposits, could act as an economic barometer.

2. **Index numbers help in policy formulation:** Many aspects of economic activity are related to price movements. The price indexes can be used as indicators of change in various segments of the economy. For example, by examining the price indexes of different segments of a firm's operations, the management can assess the impact of price changes and accordingly take some remedial and/or preventive actions. In the same way, by examining the population index, the government can assess the need to formulate a policy for health, education, and other utilities.
3. **Index numbers reveal trends and tendencies:** An index number is defined as a relative measure describing the average change in the level of a phenomenon between the current period and a base period. This property of the index number can be used to reflect typical patterns of change in the level of a phenomenon. For example, by examining the index number of industrial production, agricultural production, imports, exports, and wholesale and retail prices for the last 8–10 years, we can draw the trend of the phenomenon under study and also draw conclusions as to how much change has taken place due to the various factors.
4. **Index numbers help to measure purchasing power:** In general, the purchasing power is not associated with a particular individual; rather it is related to an entire class or group. Furthermore, it is not associated with the cost of a single item, because individuals purchase many different items in order to live. Consequently, earnings of a group of people or class must be adjusted with a price index that provides an overall view of the purchasing power for the group.

For example, suppose a person earns Rs 1000 per month in 1990. If an item costs Rs 100 in that year, the person could purchase $1000 \div 100 = 10$ units of the item with one month's earnings. But if in year 2000, the same person earns Rs 2000 per month but the item cost is Rs 250, then he could purchase $2000 \div 250 = 8$ units of the item. Hence, the effect of monthly earning relative to the particular item is less in year 2000 than in 1990 as a lesser number of units of the items can be purchased with current earnings. By dividing the item price in both the years, we can eliminate the effect of price and determine the real purchasing power for that item. For instance, in 1990, the purchasing power was $10 \div 1000 = 0.10$ or 10 paise which it was Rs 0.125 or 12.5 paise in 2000.

5. **Index numbers help in deflating various values:** When real rupee value is computed, the base period is earlier than the given years for which this value is being determined. Thus the adjustment of current rupee value to real terms is referred to as *deflating a value series* because prices typically increase over time.

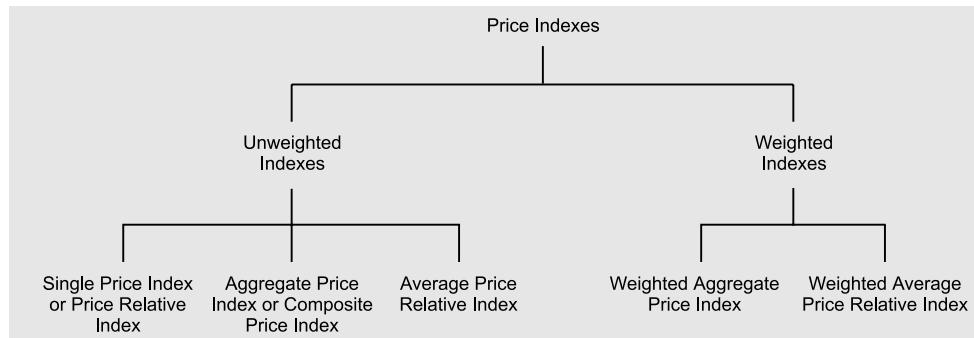
The price index number is helpful in deflating the national income to remove the effect of inflation over a long term, so that we may understand whether there is any change in the real income of the people or not. The retail price index is often used to compute real changes in earnings and expenditure as it compares the purchasing power of money at different points in time. It is generally accepted as a standard measure of inflation even though calculated from a restricted basket of goods.

Conceptual Questions 13A

1. Explain the significance of index numbers.
2. Explain the differences among the three principal types of indexes: price, quantity, and value.
3. How are index numbers constructed? What is their purpose?
4. What is an index number? Describe briefly its applications in business and industry.
5. What does an index number measure? Explain the nature and uses of index numbers.
6. Index numbers are economic barometers. Explain this statement and mention the limitations of index numbers (if any).
7. What are the basic characteristics of an index number?
8. Since value of the base year is always 100, it does not make any difference which period is selected as the base on which to construct an index. Comment.
9. What are the main uses of an index number?
10. What is meant by the term deflating a value series?

13.5 METHODS FOR CONSTRUCTION OF PRICE INDEXES

Various types of price indexes and their methods of construction can be classified into broad categories as shown in the chart below:



13.6 UNWEIGHTED PRICE INDEXES

The unweighted price indexes are further classified into three groups as shown above in the chart. The method of calculating each of these is discussed below:

13.6.1 Single Price Index

A *single unweighted price index number measures the percentage change in price for a single item or a basket of items between any two time periods.* Unweighted implies that all the values considered in calculating the index are of equal importance.

An unweighted single price index is calculated by dividing the price of an item in the given period by the price of the same item in the base period. To facilitate comparison with other years, the actual price of the item can be converted into a *price relative*, which expresses the unit price in each year (period) as a percentage of the unit price in a base year.

The general formula for calculating the single price index or price relative index is

$$\text{Single price index in period } n = \frac{p_n}{p_0} \times 100$$

where p_n = price per unit of an item in the n th year

p_0 = price per unit of an item in the base year

Example 13.1: The retail price of a typical commodity over a period of four years is given below:

Year	:	2000	2001	2002	2003
Price (Rs)	:	24.60	25.35	26.00	26.50

- (a) Find the price index based on 2000 prices
- (b) Find the percentage change in price between consecutive years (base year = 2000)
- (c) Find the percentage increase between consecutive years

Solution: (a) For the prices of the commodity with base year 2000, the price relatives for one unit of the commodity in the years 2000 to 2003 are given in Table 13.2.

Table 13.2: Price Relatives

Year	Price (Rs)	Price Relatives	Percentage Change
2000	24.60	100	—
2001	25.35	$\frac{25.35}{24.60} \times 100 = 103.04$	3.04
2002	26.00	$\frac{26}{24.60} \times 100 = 105.69$	2.65
2003	26.50	$\frac{26.50}{24.60} \times 100 = 107.72$	2.03

(c) The percentage change in price relative is divided by the index it has come from and multiplied by 100 for finding percentage increase.

$$\text{For year 2001: } \frac{103.04 - 100}{100} \times 100 = 3.04 \text{ per cent}$$

$$\text{For year 2002: } \frac{105.69 - 103.04}{103.04} \times 100 = 2.57 \text{ per cent}$$

$$\text{For year 2003: } \frac{107.72 - 105.69}{105.69} \times 100 = 1.92 \text{ per cent}$$

13.6.2 Aggregate Price Index

An **aggregate index price** or *composite price index measures the average price change for a basket of related items from the base period to the current period*. For example, to measure the change in the cost of living over a period of time, we need the index that measures the change based on the price changes for a variety of commodities including food, housing, clothing, transportation, health care, and so on. Since the number of commodities is large, therefore a sample of commodities should be selected for calculating the aggregate price index.

Irrespective of the units of measurement in which prices of several commodities are quoted, the steps of the method to calculate an aggregate price index are summarized as follows:

- Add the unit prices of a group of commodities in the year of interest.
- Add the unit prices of a group of commodities in the base year.
- Divide the sum obtained in step (i) by the sum obtained in step (ii), and multiply the quotient by 100.

From the sample of commodities or items included in the calculation of index, we cannot expect a true reflection of price changes for all commodities. This calculation provides us with only a rough estimate of price change.

A formula of calculating an unweighted aggregate price index is defined as:

$$\text{Aggregate price index } P_{01} = \frac{\sum p_1}{\sum p_0} \times 100 \quad (13-2)$$

where p_1 = unit price of a commodity in the current period of interest

p_0 = unit price for a commodity in the base period

Aggregate price index: A composite price index based on the prices of a group of commodities or items.

Example 13.2: The following are two sets of retail prices of a typical family's shopping basket. The data pertain to retail prices during 2001 and 2002.

Commodity	Unit Price (Rs)	
	2001	2002
Milk (1 litre)	18	20
Eggs (1 dozen)	15	18
Butter (1 kg)	120	150
Bread (500 gm)	9	11

Calculate the simple aggregate price index for 2002 using 2000 as the base year.

Solution: Calculations for aggregate price index are shown in Table 13.3.

Table 13.3: Calculation of Aggregate Price Index

Commodity	Unit Price (Rs)	
	2000 (p_0)	2002 (p_1)
Milk (1 litre)	18	20
Egg (1 dozen)	15	18
Butter (1 kg)	120	150
Bread (500 gm)	9	11
Total	162	199

The unweighted aggregate price index for expenses on a few food items in 2002 is given by

$$P_{01} = \frac{\sum p_1}{\sum p_0} \times 100 = \frac{199}{162} \times 100 = 122.83$$

The value $P_{01} = 122.83$ implies that the price of food items included in the price index has increased by 22.83% over the period 2000 to 2002.

Limitations of an Unweighted Aggregate Price Index

1. The unweighted aggregate approach of calculating a composite price index is heavily influenced by the items with large per unit price. Consequently items with relatively low unit price are dominated by the high unit price items.
2. Equal weights are assigned to every commodity included in the index irrespective of the relative importance of the commodity in terms of the amount purchased by a typical consumer. In other words, it did not attach more weight or importance to the price change of a high-use commodity than it did to a low-use commodity. For example, a family may purchase 30 packets of 500 gm bread in a month while it is unusual to buy 30 kg butter every month. A substantial price change for slow-moving items like butter, ghee can distort an index.

Due to these limitations, the unweighted index is not widely used in statistical analyses. These limitations suggest the use of weighted index. There are two methods to calculate weighted index, and these will be discussed later in the chapter.

13.6.3 Average Price Relative Index

This index is an improvement over the aggregate price index because it is not affected by the unit in which prices are quoted. However, it also suffers from the problem of equal importance (weight) given to all the items or commodities included in the index.

Steps of the method to calculate **average price relative** index are summarized as follows:

- (i) Select a base year, and then divide the price of each commodity in the current year by the price in the base year, to obtain price relatives.
- (ii) Divide the sum of the price relatives of all commodities by the number of commodities used in the calculation of the index.
- (iii) Multiply the average value obtained in step (ii) by 100 to express it in percentage.

Unweighted aggregate price index: A composite price index in which the price of commodities or items are weighted in accordance of their relative importance.

Price relative: A price index for a given commodity or item that is computed by dividing a current unit price by a base-period unit price and multiplying the result by 100.

The formula for computing the index is as follows:

$$\text{Average price relative index } P_{01} = \frac{1}{n} \sum \left(\frac{p_1}{p_0} \right) 100 \quad (13-3)$$

where n = number of commodities included in the calculation of the index.

The average used in computing the index of price relatives could be arithmetic mean or geometric mean. When geometric mean is used for averaging the price relatives, the formula (13-3) becomes

$$\log P_{01} = \frac{1}{n} \sum \log \left\{ \left(\frac{p_1}{p_0} \right) 100 \right\} = \frac{1}{n} \sum \log P ; \quad P = \left(\frac{p_1}{p_0} \right) 100$$

Then $P_{01} = \text{antilog} \left\{ \frac{1}{n} \sum \log p \right\}$

Advantages and Limitations of Average Price Relative Index

Advantages: This index has the following advantages over the aggregate price index:

- (i) The value of this index is not affected by the units in which prices of commodities are quoted. The price relatives are pure numbers and therefore are independent of the original units in which they are quoted.
- (ii) Equal importance is given to each commodity and extreme commodities do not influence the index number.

Limitations: Despite the few advantages mentioned above, this index is not popular on account of the following limitations.

- (i) Since it is an unweighted index, therefore each price relative is given equal importance. However in actual practice a few price relatives are more important than others.
- (ii) Although arithmetic mean is often used to calculate the average of price relatives, it also has a few biases. The use of geometric mean is computationally difficult. Other measures of central tendency such as median, mode and harmonic mean, are almost never used for calculating this index.
- (iii) Index of price relatives does not satisfy all criteria such as identity, time reversal, and circular properties, laid down for an ideal index. These criteria will be discussed later in the chapter.

Example 13.3: From the data given below, construct the index of price relatives for the year 2002 taking 2001 as base year using (a) arithmetic mean and (b) geometric mean.

Expenses on	:	Food	Rent	Clothing	Education	Misc.
Price (Rs), 2001 :		1800	1000	700	400	700
Price (Rs), 2002 :		2000	1200	900	500	1000

Solution: Calculations of Index number using arithmetic mean (A.M.) is shown in Table 13.4

Table 13.4: Calculation of Index Using A.M.

Expenses on	Price in		Price Relatives $\frac{p_1}{p_0} \times 100$
	2001 (p_0)	2000 (p_1)	
Food	1800	2000	111.11
Rent	1000	1200	120.00
Clothing	700	900	128.57
Education	400	500	125.00
Miscellaneous	700	1000	142.86
			627.54

$$\begin{aligned}\text{Average of price relative index } P_{01} &= \frac{1}{n} \sum \left(\frac{p_1}{p_0} \right) 100 \\ &= \frac{1}{5} (627.54) = 125.508\end{aligned}$$

Hence, we conclude that prices of items included in the calculation of index have increased by 25.508% in 2002 as compared to the base year 2001.

(b) Index number using geometric mean (G.M.) is shown in Table 13.5

Table 13.5: Calculations of Index Using G.M.

Expenses on	Price in	Price in	Price Relatives	Log P
	2001 (p_0)	2002 (p_2)	$P = \frac{p_1}{p_0} \times 100$	
Food	1800	2000	111.11	2.0457
Rent	1000	1200	120.00	2.0792
Clothing	700	900	128.57	2.1090
Education	400	500	125.00	2.0969
Miscellaneous	700	1000	142.86	2.1548
				10.4856

$$\begin{aligned}\text{Average price relative index } P_{01} &= \text{antilog} \left\{ \frac{1}{n} \sum \log p \right\} = \text{antilog} \left\{ \frac{1}{5} (10.4856) \right\} \\ &= \text{antilog} (2.0971) = 125.00\end{aligned}$$

Self-Practice Problems 13A

- 13.1** The following data concern monthly salaries for the different classes of employees within a small factory over a 3-year period.

Employee	Salary per Month			
	Class	1998	1999	2000
A	2300	2500	2600	
B	1900	2000	2300	
C	1700	1700	1800	
D	1000	1100	1300	

Using 1998 as the base year, calculate the simple aggregate price index for the years 1999 and 2000.

- 13.2** The following data describe the average salaries (Rs in 1000) for the employees in a company over ten consecutive years.

Year :	1	2	3	4	5
Average salary :	10.9	11.4	12.0	12.7	13.6
Year :	6	7	8	9	10
Average salary :	14.4	15.0	15.5	16.3	17.6

- (a) Calculate an index for these average salaries using year 5 as the base year.

- (b) Calculate the percentage points change between consecutive years.

- 13.3** A state Govt. had compiled the information shown below regarding the price of the three essential commodities: wheat, rice, and sugar. From the commodities listed, the corresponding price indicates the average price for that year. Using 1998 as the base year, express the price for the years 2000 to 2002, in terms of unweighted aggregate index.

Commodity	1998	1999	2000	2001	2002
Wheat	4	6	8	10	12
Rice	16	20	24	30	36
Sugar	8	10	16	20	24

- 13.4** Following are the prices of commodities in 2003 and 2004. Calculate a price index based on price relatives, using the geometric mean.

Year	Commodity					
	A	B	C	D	E	F
2003	45	60	20	50	85	120
2004	55	70	30	75	90	130

- 13.5** A textile worker in the city of Mumbai earns Rs 3500 per month. The cost of living index for a particular month is given as 136. Using the following information, find out the amount of money he spent on house rent and clothing.

Group	Expenditure (Rs)	Group Index
Food	1400	180
Clothing	x	150
House rent	y	100
Food and lighting	560	110
Misc.	630	80

[Delhi Univ., BCom, 1997]

- 13.6** In 1996, for working class people, wheat was selling at an average price of Rs 160 per 10 kg, cloth at Rs 40 per metre, house rent Rs 10,000 per house, and other items at Rs 100 per unit. By 1997 the cost of wheat rose by Rs 40 per 10 kg, house rent by Rs 1500 per house, and other items doubled in price. The working class cost of living index for the year 1997 (with 1996 as base) was 160. By how much did the cloth price rise during the period 1996–97?

- 13.7** From the following data calculate an index number using family budget method for the year 1996 with 1995 as the base year.

Commodity	Quantity (in units) in 1995	Price (in Rs) per unit	
		1995	1996
A	110	8.00	12.00
B	25	6.00	7.50
C	10	5.00	5.25
D	20	48.00	60.00
E	25	15.00	16.50
F	30	9.00	27.00

[Karnataka Univ., BCom, 1997]

- 13.8** The following table gives the annual income of a teacher and the general index of price during 1990–97. Prepare the index number to show the change in the real income of the teacher and comment on price increase:

Year	Income	Index
1990	4000	100
1991	4400	130
1992	4800	160
1993	5200	220
1994	5600	270
1995	6000	330
1996	6400	400
1997	6800	490

[HP Univ., BCom, 1997]

Hints and Answers

- 13.1** Simple aggregate price index

$$P_{0,89} = \frac{7300}{6900} \times 100 = 105.8 \text{ for the year 1999}$$

$$P_{0,90} = \frac{8000}{6900} \times 100 = 115.9 \text{ for the year 2000}$$

- 13.2** (a)

Year	:	1	2	3	4	5
Index number :		80.1	83.8	88.2	93.4	100
Year	:	6	7	8	9	10
Index number :		105.9	110.3	114.0	119.9	129.4

For example, index for year 1: $(10.9 \div 13.6)100 = 80.1$; year 2: $(11.4 \div 13.6)100 = 83.8$

(b)

Year	Index number	Percentage point change
1	80.1	—
2	83.8	3.7
3	88.2	4.4
4	93.4	5.2
5	100.0	6.6
6	105.9	5.9
7	110.3	4.4
8	114.0	3.7
9	119.9	5.9
10	129.4	9.5

- 13.3** Aggregate price

1998	1999	2000	2001	2002
100	133.33	137.78	125	120

13.4

Commodity	$P = \frac{P_1}{P_0} \times 100$	Log P
A	122.22	2.0872
B	116.67	2.0669
C	150.00	2.1761
D	150.00	2.1761
E	105.88	2.0248
F	108.33	2.0348

$$P_{01} = \text{antilog} \left\{ \frac{1}{n} \log P \right\} = \text{antilog} \left\{ \frac{1}{6} (12.5659) \right\}$$

$$= \text{antilog} (2.0948) = 124.4$$

- 13.5 Let expenditure on clothing be x and on house rent be y . Then as per conditions given, we have

$$3500 = 1400 + x + y + 560 + 630$$

$$\text{or } x + y = 910 \quad (\text{i})$$

Multiplying expenditure with group index and equating it to 136, we get

$$136 = \frac{(1400 \times 180) + (x \times 150) + (y \times 100)}{3500} + (500 \times 110) + (630 \times 80)$$

$$136 = \frac{2,52,000 + 150x + 100y + 61,600 + 50,400}{3500}$$

$$4,76,000 = 2,52,000 + 150x + 100y + 61,600 + 50,400$$

$$150x + 100y = 1,12,000 \quad (\text{ii})$$

Multiplying Eqn. (i) by 150 and subtracting it from (ii), we get

$$50y = 24,500 \text{ or } y = \text{Rs } 490 \text{ (house rent)}$$

Substituting the value of y in Eqn. (i): $x + 490 = 910$ or $x = \text{Rs } 420$ (clothing)

- 13.6 Let the rise in price of cloth be x .

Commodity	Price	Index	Price 1997	Index
Wheat	160	100	200	$\frac{200}{160} \times 100 = 125$
Cloth	40	100	x	$\frac{x}{40} \times 100 = 2.5x$
House rent	10,000	100	11,500	$\frac{11,500}{10,000} \times 100 = 115$
Miscellaneous	100	100	200	$\frac{200}{100} \times 100 = 200$
Total				$440 + 2.5x$

The index for 1997 as given is 160. Therefore, the sum of the index numbers of the four commodities would be $160 \times 4 = 640$. Thus $440 + 2.5x = 640$ or $x = 80$. Hence the rise in the price of cloth was Rs 40 (80 – 40) per metre.

13.7

Commodity	Quantity	p_0	p_1	$P = \frac{p_1}{p_0} \times 100$	PQ
	Q				
A	100	8	12.00	150	15,000
B	25	6	7.50	125	3,125
C	10	5	5.25	105	1,050
D	20	48	60.00	125	2,500
E	25	15	16.50	110	2,750
F	30	9	27.00	300	9,000
Total	210				33,425

$$\text{Index number} = \frac{\Sigma PQ}{\Sigma Q} = \frac{33,425}{210} = 159.17$$

13.8

Year	Income	Index	Real Income	Real Income
	(Rs)		(Rs)	Index
1990	4000	100	$\frac{4000}{100} \times 100 = 4000.00$	100.00
1991	4400	130	$\frac{4400}{130} \times 100 = 3384.62$	84.62
1992	4800	160	$\frac{4800}{160} \times 100 = 3000.00$	75.00
1993	5200	220	$\frac{5200}{220} \times 100 = 2363.64$	59.09
1994	5600	270	$\frac{5600}{270} \times 100 = 2074.07$	51.85
1995	6000	330	$\frac{6000}{\sum p_0 q_1} \times 100 = 1818.18$	45.45
1996	6400	400	$\frac{6400}{400} \times 100 = 1600.00$	40.00
1997	6800	490	$\frac{6800}{490} \times 100 = 1387.76$	34.69

13.7 WEIGHTED PRICE INDEXES

While constructing weighted price indexes, rational weights are assigned to all items or commodities in an explicit manner. Such weights indicate the relative importance of items or commodities included in the calculation of an index. The weights used are of two types, *quantity weights* and *value weights*. There are two price indexes that are commonly in use

1. Weighted aggregate price index
2. Weighted average of price relative index

13.7.1 Weighted Aggregate Price Index

In a weighted aggregate price index, each item in the basket of items chosen for calculation of the index is assigned a weight according to its importance. In most cases, the quantity of usage is the best measure of importance. Hence, we should obtain a measure of the quantity of usage for the various items in the group. This explicit weighting allows us to gather more information than just the change in price over a period of time as well as improve the accuracy of the general price level estimate.

Weight is assigned to each item in the basket in various ways and the weighted aggregates are also used in different ways to calculate an index. A few methods (or approaches) to determine weights (value) to be assigned to each item in the basket are as follows:

- Laspeyre's method
- Paasche's method
- Dorbish and Bowley's method
- Fisher's ideal method
- Marshall-Edgeworth's method
- Walsch's method
- Kelly's method

Laspeyre's Weighting Method

This method suggests to treat quantities as constant at *base period* level and are used for weighting price of each item or commodities both in base period and current period. Since this index number depends upon the same base price and quantity, therefore one can directly compare the index of one period with another. The formula for calculating *Laspeyre's price index*, named after the statistician Laspeyre's is given by

$$\text{Laspeyre's price index} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

where p_1 = prices in the current period

p_0 = prices in the base period

q_0 = quantities consumed in the base period

Advantages and Disadvantages of Laspeyre's Method

Advantages: The main advantage of this method is that it uses only one quantity measure based on the base period and therefore we need not keep record of quantity consumed in each period. Moreover, having used the same base period quantity, we can compare the index of one period with another directly.

Disadvantages: We know that the consumption of commodities decreases with relatively large increases in price and vice versa. Since in this index the fixed quantity weights are determined from the base period usage, it does not adjust such changes in consumption and therefore tends to result in a bias in the value of the composite price index.

Example 13.4: Compute the cost of living index number using Laspeyre's method, from the following information:

Commodity	Unit Consumption in Base Period	Price in Base Period	Price in Current Period
Wheat	200	1.0	1.2
Rice	50	3.0	3.5
Pulses	50	4.0	5.0
Ghee	20	20.0	30.0
Sugar	40	2.5	5.0
Oil	50	10.0	15.0
Fuel	60	2.0	2.5
Clothing	40	15.0	18.0

Laspeyre's index: A weighted aggregate price index in which the weight for each commodity or item is its base-period quantity.

Solution: Calculation of cost of living index by Laspeyre's method is shown in Table 13.6.

Table 13.6: Laspeyre's Method

Commodity	Base Period Quantity (q_0)	Base Period Price (p_0)	Current Price (p_1)	$p_1 q_0$	$p_0 q_0$
Wheat	200	1.0	1.2	240	200
Rice	50	3.0	3.5	175	150
Pulses	50	4.0	5.0	250	200
Ghee	20	20.0	30.0	600	400
Sugar	40	2.5	5.0	200	100
Oil	50	10.0	15.0	750	500
Fuel	60	2.0	2.5	150	120
Clothing	40	15.0	18.0	720	600
Total	510			3085	2270

$$\text{Cost of living index} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{3085}{2270} \times 100 = 135.9$$

Paasche's index: A weighted aggregate price index in which the weight for each commodity or item is its current-period quantity.

Paasche's Weighting Method

In the Paasche's method, the price of each item or commodity is weighted by the quantity in the current period instead of the base year as used in Laspeyre's method. Paasche's formula for calculating the index is given by

$$\text{Paasche price index} = \frac{\sum p_1 q_1}{\sum p_0 q_1}$$

where p_1 = prices in current year

p_0 = prices in base year

q_1 = quantities in current year

Advantages and Disadvantages of the Paasche's Method

Advantages: The Paasche's method combines the effects of changes in price and quantity consumption patterns during the current year. It provides a better estimate of changes in the economy than Laspeyre's method. If the prices or quantities of all commodities or items change in the same ratio, then the values of the Laspeyre's and Paasche's indexes will be same.

Disadvantages: This method requires knowledge of the quantities consumed of all commodities in each period. Getting the data on the quantities for each period is either expensive or time-consuming. Moreover, each year the index number for the previous year requires recomputation to reflect the effect of the new quantity weights. Thus, it is difficult to compare indexes of different periods when calculated by the Paasche's method.

Example 13.5: For the following data, calculate the price index number of 1999 with 1998 as the base year, using: (a) Laspeyre's method, and (b) Paasche's method.

Commodity	1998		1999	
	Price	Quantity	Price	Quantity
A	20	8	40	6
B	50	10	60	5
C	40	15	50	15
D	20	20	20	25

[Kurukshetra Univ., MBA, 1999]

Solution: Table 13.7 presents the information necessary for both Laspeyre's and Paasche's methods.

Table 13.7: Calculation of Laspeyre's and Paasche's and Paasche's Indexes

Commodity	Base Period, 1998		Current year, 1999				
	Price (p_0)	Quantity (q_0)	Price (p_1)	Quantity (q_1)	$p_1 q_0$	$p_0 q_0$	$p_1 q_1$
A	20	8	40	6	320	160	240
B	50	10	60	5	600	500	300
C	40	15	50	15	750	600	750
D	20	20	20	25	400	400	500
					2070	1660	1790
							1470

$$\text{Laspeyre's price index} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{2070}{1660} \times 100 = 124.7$$

$$\text{Paasche's price index} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{1790}{1470} \times 100 = 121.77$$

The Paasche's price index shows a price level increase of 21.77 per cent, while Laspeyre's index shows a price level increase of 24.7 per cent. Hence, we may conclude that Paasche's index shows a trend towards less expensive commodities.

Dorbish and Bowley's Method

This method (or approach) is the simple *arithmetic mean* of the Laspeyre's and Paasche's indexes. This index takes into account the influence of quantity weights of both base period and current period. The formula for calculating the index using Dorbish and Bowley method is given by

$$\text{Dorbish and Bowley's price index} = \frac{1}{2} \left\{ \frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1} \right\} 100$$

Fisher's Ideal Method

This method (or approach) is the *geometric mean* of the Laspeyre's and Paasche's indexes and the formula in given by

$$\text{Fisher's ideal price index} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

Advantages and Disadvantages of Fisher's Method

Advantages: Fisher's method is also called ideal method due to following reasons:

- (i) The formula is based on geometric mean which is considered to be the best average for constructing index numbers.
- (ii) The formula takes into account both base year and current year quantities as weights. Thus it avoids the bias associated with the Laspeyre's and Paasche's indexes.
- (iii) This method satisfies essential tests required for a index, that is, time reversal test and factor reversal test.

Disadvantages: The calculation of index using this method requires more computation time. Although the index number is theoretically better than others discussed previously, it is not fit for common use because it requires current quantity weights every time an index is calculated.

Example 13.6: Compute index number from the following data using Fisher's ideal index formula.

Commodity	1999		2000	
	Price	Quantity	Price	Quantity
A	12	10	15	12
B	15	7	20	5
C	24	5	20	9
D	5	16	5	14

Solution: Table 13.8 presents the information necessary for Fisher's method to calculate the index.

Table 13.8: Calculations of Fisher Ideal Index

Commodity	Base Year, 1999		Current Year, 2000		$p_1 q_0$	$p_0 q_0$	$p_1 q_1$	$p_0 q_1$
	(q_0)	(p_0)	(q_1)	(p_1)				
A	12	10	15	12	144	120	180	150
B	15	7	20	5	75	105	100	140
C	24	5	20	9	216	120	180	100
D	5	16	5	14	70	80	70	80
					505	425	530	470

$$\text{Fisher's ideal price index} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 = \sqrt{\frac{505}{425} \times \frac{530}{470}} \times 100 \\ = \sqrt{1.3399} \times 100 = 1.1576 \times 100 = 115.76$$

Hence, we conclude that the price level has increased by 15.76% in the year 2000.

Example 13.7: Calculate from the following data, the Fisher's ideal index number for the year 2000:

Commodity	1999		2000	
	Price (Rs)	Expenditure on Quantity Consumed (Rs)	Price (Rs)	Expenditure on Quantity Consumed (Rs)
A	8	200	65	1950
B	20	1400	30	1650
C	5	80	20	900
D	10	360	15	300
E	27	2160	10	600

Solution: Table 13.9 presents the information necessary for Fisher's method to calculate the index.

Table 13.9: Calculations of Fisher's Ideal Index

Commodity	Base Year, 1999		Current Year, 2000		$p_1 q_0$	$p_0 q_0$	$p_1 q_1$	$p_0 q_1$
	(p_0)	(q_0)	(p_1)	(q_1)				
A	8	$200/8 = 25$	65	$1950/65 = 30$	200	1950	240	240
B	20	$1400/20 = 70$	30	$1650/30 = 55$	2100	1400	1650	1100
C	5	$80/5 = 16$	20	$900/20 = 45$	320	80	900	225
D	10	$360/10 = 36$	15	$300/15 = 20$	540	360	300	200
E	27	$2160/27 = 80$	10	$600/10 = 60$	800	2160	600	1620
					5385	4200	5400	3385

$$\text{Fisher's ideal price index} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 = \sqrt{\frac{5385}{4200} \times \frac{5400}{3385}} \times 100 \\ = 1.430 \times 100 = 143.$$

Hence we conclude that the price level has increased by 43% in the year 2000.

Marshall-Edgeworth Method

In this method the sum of base year and current year quantities are considered as the weight to calculate the index. The formula for constructing the index is:

$$\text{Marshall-Edgeworth price index} = \frac{\sum(q_0 + q_1)p_1}{\sum(q_0 + q_1)p_0} \times 100 = \frac{\sum q_0 p_1 + \sum q_1 p_1}{\sum q_0 p_0 + \sum q_1 p_0} \times 100$$

where notations have their usual meaning.

The disadvantage with this formula is the same as that of Paasche index and Fisher's ideal index in the sense that it also needs current quantity weights every time an index is constructed.

Walsch's Method

In this method the quantity weight used is the geometric mean of the base and current year quantities. The formula for constructing the index is

$$\text{Walsch's price index} = \frac{\sum p_1 \sqrt{q_0 q_1}}{\sum p_0 \sqrt{q_0 q_1}} \times 100$$

Although this index satisfies the time reversal test, it needs current quantity weight every time an index is constructed.

Kelly's Method

The method suggested by T L Kelly for the construction of index number is

$$\text{Kelly's price index} = \frac{\sum p_1 q}{\sum p_0 q} \times 100$$

where q = fixed weight.

This method is also called the *fixed weight aggregate method* because instead of using base period or current period quantities as weights, it uses weights from a representative period. The representative weights are referred to as *fixed weight*. The fixed weights and the base period prices do not have to come from the same period.

Advantages and Disadvantages of Kelly's Method

Advantages: An important advantage of this index is that it does not need yearly changes in the weights. One can select a different period for fixed weight other than base period. This can improve the accuracy of the index. Moreover, the base period can also be changed without changing the fixed weight. The weights should be appropriate and should indicate the relative importance of various commodities. This weight may be kept fixed until new data are available to revise the index.

Disadvantages: One disadvantage with this index is that it does not take into account the weight either of the base year or of the current year.

Example 13.8: It is stated that the Marshall-Edgeworth index number is a good approximation of the ideal index number. Verify this statement using the following data:

Commodity	2002		2003	
	Price	Quantity	Price	Quantity
A	2	74	3	82
B	5	125	4	140
C	7	40	6	33

Solution: Table 13.10 presents the information necessary to calculate Fisher and Marshall-Edgeworth indexes.

Table 13.10: Calculations of Fisher's Ideal and Marshall-Edgeworth's Index

Commodity	Base Year, 2002		Current Year, 2003		$p_1 q_0$	$p_0 q_0$	$p_1 q_1$	$p_0 q_1$
	(p_0)	(q_0)	(p_1)	(q_1)				
A	2	74	3	82	222	148	246	164
B	5	125	4	140	500	625	560	700
C	7	40	6	33	240	280	198	231
					962	1053	1004	1095

$$\begin{aligned}
 \text{Fisher ideal price index} &= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 = \sqrt{\frac{962}{1053} \times \frac{1004}{1095}} \times 100 \\
 &= \sqrt{0.836} \times 100 = 0.9144 \times 100 = 91.44 \\
 \text{Marshall-Edgeworth price index} &= \frac{\sum p_1 (q_0 + q_1)}{\sum p_0 (q_0 + q_1)} \times 100 = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100 \\
 &= \frac{962 + 1004}{1053 + 1095} \times 100 = 0.9152 \times 100 = 91.52
 \end{aligned}$$

Hence, we conclude that Fisher's method and Marshall-Edgeworth method provide almost the same value of the index.

Example 13.9: Compute Laspeyre's, Paasche's, Fisher's, and Marshall-Edgeworth's index numbers from the following data:

Item	1998		1999	
	Price	Quantity	Price	Quantity
A	5	25	6	30
B	3	8	4	10
C	2	10	3	8
D	10	4	3	5

[Bangalore Univ., BCom, 2000]

Solution: Table 13.11 presents the information necessary to calculate several indexes.

Table 13.11: Calculations of Indexes

Item	Base Year, 1998		Current Year, 1999		$p_1 q_0$	$p_0 q_0$	$p_1 q_1$	$p_0 q_1$
	(p_0)	(q_0)	(p_1)	(q_1)				
A	5	25	6	30	150	125	180	150
B	3	8	4	10	32	24	40	30
C	2	10	3	8	30	20	24	16
D	10	4	3	5	12	40	15	50
					224	209	259	246

$$\text{Laspeyre's price index} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{224}{209} \times 100 = 107.17$$

$$\text{Paasche's price index} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{259}{246} \times 100 = 105.28$$

$$\text{Fisher's ideal price index} = \sqrt{\sum p_1 q_0 \sum p_0 q_1} = \sqrt{107.17 \times 105.28} = 106.22$$

$$\begin{aligned}
 \text{Marshall-Edgeworth's price index} &= \frac{\sum p_1 (q_0 + q_1)}{\sum p_0 (q_0 + q_1)} \times 100 = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100 \\
 &= \frac{244 + 259}{209 + 246} \times 100 = 110.55
 \end{aligned}$$

13.7.2 Weighted Average of Price Relative Index

Unlike the unweighted average of price relative, the weighted average of price relative is determined by using the quantity consumed in the base period for weighting the items or commodities. The value (in rupees) of each item or commodity included in the calculation of composite index is determined by multiplying the price of each item by its quantity consumed.

The formula for constructing the weight average of price relatives index using base values is:

$$\begin{aligned}\text{Weighted average of price relative index, } P_{01} &= \frac{\sum \{(p_1/p_0) \times 100\}(p_0 q_0)}{\sum p_0 q_0} = \frac{\Sigma PV}{\Sigma V} \\ &= \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100\end{aligned}$$

where $V (= p_0 q_0)$ = base period value

$P (= (p_1/q_0) \times 100)$ = price relative

This formula is equivalent to Laspeyre's method for any given problem.

If we wish to compute a weighted average of price relative using $V = p_0 q_1$, then the above formula becomes

$$P_{01} = \frac{\sum \{(p_1/p_0) \times 100\}(p_0 q_1)}{\sum p_0 q_1} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

This formula is equivalent to Paasche's method for any given problem.

If instead of using weighted arithmetic average, we wish to use weighted geometric mean, then the above formula becomes

$$P_{01} = \frac{\Sigma V \times \log P}{\Sigma V}; \quad P = \frac{p_1}{p_0} \times 100 \text{ and } V = p_0 q_0$$

Advantages of Weighted Average Price Relatives

- (i) Different index numbers constructed using average price relative with same base can be combined to form a new index.
- (ii) Weighted average of price relative method is suitable to construct an index by selecting one item from each of the many subgroups of items. In such a case, the values of each subgroup may be used as weights.

Example 13.10: A large manufacturer purchases an identical component from three different suppliers that differ in unit price and quantity supplied. The relevant data for 2000 and 2001 are given below:

Supplier	Quantity Index in (2000)		Unit Price (Rs)	
			2000	2001
A	20		18	20
B	40		12	14
C	10		15	16

Construct a weighted average price relative index using (a) arithmetic mean and (b) geometric mean.

Solution: Table 13.12 presents the information necessary to calculate the weight average price relative index.

Table 13.12: Calculations of Weighted Average of Price Relatives

Supplier	Prices in		Quantity in 2000	Percentage Price Relative	Base Value $V = p_0 q_0$	Weighted Percentage Relative PV
	2000	2001				
	p_0	p_1	q_0	$P = \frac{p_1}{p_0} \times 100$		
A	18	20	20	$(20/18) \times 100 = 111.11$	360	39,999.60
B	12	14	40	$(14/12) \times 100 = 116.67$	480	56,001.60
C	15	16	10	$(16/15) \times 100 = 106.67$	150	16,000.50
					990	1,12,001.70

(a) Weighted average of price relative index

$$P_{01} = \frac{[\Sigma(p_1/p_0)100] p_0 q_0}{\Sigma p_0 q_0} = \frac{1,12,001.70}{990} = 113.13$$

The value of P_{01} implies that there has been 13.13% increase in price from year 2000 to 2001.

(b)

Table 13.13: Calculations of Weighted Geometric Mean of Price Relatives

Supplier	Prices in 2000		Quantity in 2000	Base Value	Percentage Price Relative $P = \frac{p_1}{p_0} \times 100$	Log P	$V \log P$
	p_0	p_1					
A	18	20	20	360	111.11	2.046	736.56
B	12	14	40	480	116.67	2.067	992.16
C	15	16	10	150	106.67	2.028	304.20
				990			2032.92

Weighted geometric mean of price relatives

$$P_{01} = \text{antilog} \left\{ \frac{\sum V \times \log P}{\sum V} \right\} = \text{antilog} \left\{ \frac{2032.92}{990} \right\} \\ = \text{antilog} (2.0535) = 113.11$$

13.8 QUANTITY OR VOLUME INDEXES

A quantity index measures the percentage change in consumption, production or distribution level of either an individual item or a basket of items from one time period to another. When constructing quantity indexes, it is necessary to *hold price levels constant over time* to isolate the effect of quantity (consumption level) changes only. For example, agricultural production is measured using a quantity index because it eliminates effects of fluctuating prices. Any of the methods, such as the relative method (both simple and weighted) or the aggregative method, which take into account weights to construct price indexes can also be used to calculate quantity indexes. The weights in quantity index numbers are prices. Therefore quantity indexes can be easily derived from a price indexes by interchanging the p 's and q 's.

Seven quantity indexes analogous to the seven price indexes already discussed in the previous section can be constructed as given below:

$$\text{Laspeyres quantity index } Q_L = \frac{\sum V_0 (q_1/q_0)}{\sum V_0} \times 100 = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$$

where $V_0 = p_0 q_0$, values of base year consumption at base year prices

$$\text{Similarly Paasche's quantity index } Q_P = \frac{\sum V_1 (q_1/q_0)}{\sum V_1} \times 100 = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$$

where $V_1 = p_0 q_0$, values of base year consumption at current year prices

$$\text{Fisher's quantity index } Q_F = \sqrt{Q_L \times Q_P} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_1} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100$$

The formula for computing a weighted average of quantity relative index is also the same as used to compute a price index. The formula for this type of quantity index is

$$\text{Weighted average of quantity relative index} = \frac{\sum \left(\frac{q_1}{q_0} \times 100 \right) (q_0 p_0)}{\sum q_0 p_0}$$

where q_1 = quantities for the current period
 q_0 = quantities for the base period

Example 13.11: Obtain Laspeyre's price index number and Paasche's quantity index number from the following data:

Item	Price (Rs per Unit)		Quantity (Units)	
	Base Year	Current Year	Base Year	Current Year
1	2	5	20	15
2	4	8	4	5
3	1	2	10	12
4	5	10	5	6

[Mangalore Univ., BCom, 1997]

Solution: Table 13.14 presents the information necessary to calculate Laspeyre's price index and Paasche's quantity indexes.

Table 13.14: Calculations on Laspeyre's Price Index and Paasche's Quantity Index

Item	Price		Quantity		$p_1 q_0$	$p_0 q_0$	$q_1 p_1$	$q_0 p_1$
	p_0	p_1	q_0	q_1				
1	2	5	20	15	100	40	75	100
2	4	8	4	5	32	16	40	32
3	1	2	10	12	20	10	24	20
4	5	10	5	6	50	25	60	50
					202	91	199	202

$$\text{Laspeyre's price index} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{202}{91} \times 100 = 221.98$$

$$\text{Paasche's quantity index} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100 = \frac{199}{202} \times 100 = 98.51$$

Example 13.12: Compute the quantity index by using Fisher's formula from the data given below:

Commodity	2002		2003	
	Price (Rs/Unit)	Total Value	Price (Rs/Unit)	Total Value
A	5	50	4	48
B	8	48	7	49
C	6	18	5	20

Solution: The base year quantity q_0 and current year quantity q_1 for individual commodity can be calculated as follows:

$$q_0 \text{ (for 2002)} = \frac{\text{Total value}}{\text{Price}} = \frac{50}{5} = 10; \quad \frac{48}{4} = 6; \quad \frac{18}{6} = 3$$

$$q_1 \text{ (for 2003)} = \frac{\text{Total value}}{\text{Price}} = \frac{48}{4} = 12; \quad \frac{49}{7} = 7; \quad \frac{20}{5} = 4$$

Table 13.15: Calculations for Fisher's Quantity Index

Commodity	Price, 2002		Quantity, 2003		$p_1 q_0$	$p_0 q_0$	$q_1 p_1$	$q_0 p_1$
	p_0	p_1	q_0	q_1				
A	5	10	4	12	60	50	48	40
B	8	6	7	7	56	48	49	42
C	6	3	5	4	24	18	20	15
					140	116	117	97

Substituting values in the formula, we get

$$\begin{aligned}\text{Fisher's quantity index} &= \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100 \\ &= \sqrt{\frac{140}{116} \times \frac{117}{97}} \times 100 = 120.65\end{aligned}$$

Example 13.13: Calculate the weighted average of quantity relative index from the following data:

Commodity	Quantity (Units)		Price (Rs/Unit) 2000
	2000	2002	
A	10	12	100
B	15	20	75
C	8	10	80
D	20	25	60
E	50	60	500

Solution: Table 13.16 presents the information necessary to calculate the weighted average of quantity relative index.

Table 13.16: Calculations of a Weighted Average of Quantity Relatives Index

Commodity	Quantity (units)		Price (Rs/unit) 2000 p_0	Percentage Relatives $(q_1/q_0) \times 100$	Base Value $q_0 p_0$	Weighted Relatives $\{(q_1/q_0) \times 100\} \times q_0 p_0$
	2000	2002				
	q_0	q_1				
A	10	12	100	$(12/10) \times 100 = 120$	1000	1,20,000.00
B	15	20	75	$(20/15) \times 100 = 133.33$	1125	1,49,996.25
C	8	10	80	$(10/8) \times 100 = 125$	640	80,000.00
D	20	25	60	$(25/20) \times 100 = 125$	1200	1,50,000.00
E	50	60	500	$(60/50) \times 100 = 120$	25,000	30,00,000.00
					28,965	34,99,996.25

$$\begin{aligned}\text{Weighted average of quantity relatives index} &= \frac{\sum \{(q_1/q_0) \times 100\} (q_0 p_0)}{\sum q_0 p_0} \\ &= \frac{34,99,996.25}{28,965} = 120.835\end{aligned}$$

13.9 VALUE INDEXES

A value index number measures the percentage change in the total value of either an individual item or a basket of items from one time period to another. The value of an item or commodity is obtained by multiplying its price and quantity. Since value is determined both by price and quantity, a value index measures the combined effects of price and quantity changes. A simple value ratio is equal to the value of the current year divided by the value of the base year. If this ratio is multiplied by 100, we get the value index as:

$$\text{Value index, } V = \frac{\sum p_1 q_1}{\sum p_0 q_0} \times 100$$

If the values are given directly, then the value index number is given by

$$\text{Value index, } V = \frac{\sum V_1}{\sum V_0}$$

where V_0 = values at the base year or period

V_1 = values at the current year or period

Such indexes are not weighted because they take into account both the price and quantity. These indexes are, however, not very popular because the situation revealed by price and quantities are not fully revealed by the values. A value index does not distinguish between the effects of its components, namely price and quantity.

Self-Practice Problems 13B

- 13.9** The following table contains information from the raw material purchase records of a small factory for the year 2002 and 2003:

Commodity	2002		2003	
	Price (Rs/Unit)	Total Value	Price (Rs/Unit)	Total Value
A	5	50	6	72
B	7	84	10	80
C	10	80	12	96
D	4	20	5	30
E	8	56	8	64

Calculate Fisher's ideal index number.

- 13.10** The subgroup indexes of the consumer price index number for urban non-manual employees of an industrial centre for a particular year (with base 1990 = 100) were:

Food	200
Clothing	130
Fuel and Lighting	120
Rent	150
Miscellaneous	140

The weights are 60, 8, 7, 10, and 15 respectively. It is proposed to fix dearness allowance in such a way as to compensate fully the rise in the prices of food and house rent. What should be the dearness allowance, expressed as a percentage of wage?

[Kurukshetra Univ., MBA, 1996]

- 13.11** The owner of a small shop selling food items collected the following information regarding the price and quantity sold of a particular item.

Item	Average Price (Rs/Unit)		Quantity Sold (Units)	
	2002	2003	2002	2003
A	1	2	10	5
B	1	x	5	2

If the ratio between Laspeyre's (L) and Paasche's (P) Index number is: L:P = 28:27, then find the value of x.

- 13.12** An increase of 50 per cent in the cost of a certain consumable product raises the cost of living of a certain family by 5 per cent. What percentage of its cost of living was due to buying that product before the change in the price?

- 13.13** Calculate Fisher's ideal index from the data given below:

Commodity	Base Year, 2000		Current Year, 2001	
	Price	Value	Price	Value
A	10	30	12	48
B	15	60	15	75
C	5	50	8	96
D	2	10	3	25

[HP Univ., MCom, 1995]

- 13.14** Using the data given below, calculate the price index number for the year 1998 by (i) Laspeyre's formula, (ii) Paasche's formula, and (iii) Fisher's formula considering 1989 as the base year.

Commodity	Price (Rs/Unit)		Quantity (in 1000 kg)	
	1989	1998	1989	1998
Rice	9.3	4.5	100	90
Wheat	6.4	3.7	11	10
Pulses	5.1	2.7	5	3

- 13.15** It is stated that the Marshall-Edgeworth's index is a good approximation of the ideal index number. Verify using the following data:

Commodity	1996		2000	
	Price	Quantity	Price	Quantity
A	2	74	3	82
B	5	125	4	140
C	7	40	6	33

- 13.16** In preparation for an appropriations hearing, the DCP of a city zone has collected the following information:

Type of Crime	2000	2001	Weight
Robberies	13	8	6
Car thefts	15	22	5
Cycle thefts	249	185	4
Pocket picking	328	259	1
Theft by servants	497	448	2

Calculate the index of crime for 2001, using 2000 as the base period.

- 13.17** Using Paasche's formula compute the quantity index for the year 1993 with 1985 as base year.

Commodity	Quantity (in Units)		Value (in Rs)	
	1985	1993	1985	1992
A	100	150	500	900
B	80	100	320	500
C	60	72	150	360
D	30	33	360	297

- 13.18** Calculate a weighted average of relative quantity index using 1995 as base period:

Commodity	Quantity (in 1000 kg)		Price (Rs/kg)
	1995	1999	
Wheat	29	24	3.80
Corn	3	2.5	2.91
Soyabean	12	14	6.50

Hints and Answers

- 13.9** Divide the values by price and obtain quantity figures and then calculate Fisher's ideal price index.

Commodity	p_0	q_0	p_1	q_1	$p_1 q_0$	$p_0 q_0$	$q_1 p_1$	$q_0 p_1$
A	5	10	6	12	60	50	72	60
B	7	12	10	8	120	84	80	56
C	10	8	12	8	96	80	96	80
D	4	5	5	6	25	20	30	24
E	8	7	8	8	56	56	64	64
					357	290	342	284

Fisher's ideal price index:

$$\sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 = \sqrt{\frac{357}{290} \times \frac{342}{284}} \times 100 = 121.96$$

- 13.10** Let the income of the consumer be Rs 100. He spent Rs 60 on food and Rs 10 on house rent in 1990. The index of food is 200 and the house rent Rs 150 for the particular year for which the data are given. In order to maintain the same consumption standards regarding two items, he will have to spend Rs 120 on food and Rs 15 on house rent. Further the weights of other items are constant; in order to maintain the same standard he will have to spend $120+8+7+15+5 = \text{Rs } 155$. Hence the dearness allowance should be 55 per cent.

13.11

Item	p_0	q_0	p_1	q_1	$p_1 q_0$	$p_0 q_0$	$p_1 q_1$	$p_0 q_1$
A	1	10	2	5	20	10	10	5
B	1	5	x	2	$5x$	5	$2x$	2
					$20 + 5x$	15	$10 + 2x$	7

$$\text{Laspeyre's index} = \frac{\sum p_1 q_1}{\sum p_0 q_0} = \frac{20 + 5x}{15};$$

$$\text{Paasche's index} = \frac{\sum p_1 p_0}{\sum p_0 q_1} = \frac{10 + 2x}{7}$$

$$\text{Given } \frac{(20 + 5x)/15}{(10 + 2x)/7} = \frac{28}{27}$$

$$\text{or } \frac{20 + 5x}{15} \times \frac{7}{10 + 2x} = \frac{28}{27} \text{ or } x = 4$$

- 13.12** Let the cost of the article before the increase be x . After increase it will be $150x/100 = 1.5x$. The rise $1.5x - x = 0.5x$ is equivalent to an increase of 5 per cent in the cost of living. The increases in the cost of living was $1.05y - y = 0.05y$.

Hence $0.5x = 0.05y$ or $x = 0.5y/0.5 = 0.1y = 10$ per cent of y . Thus the expenditure on that item was 10 per cent of the cost of living.

13.13

Commodity	p_0	q_0	p_1	q_1	$p_1 q_0$	$p_0 q_0$	$p_1 q_1$	$p_0 q_1$
A	10	3	12	4	36	30	48	40
B	15	4	15	5	60	60	75	75
C	5	10	8	12	50	50	96	60
D	2	5	3	8	10	10	15	10
					191	150	234	185

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 = \sqrt{\frac{191}{150} \times \frac{234}{185}} \times 100 = 126.9$$

13.14

Commodity	Price (Rs)		Quantity	
	1989	1998	1989	1998
	p_0	q_0	q_1	$p_1 q_0$
Rice	9.3	4.5	100	90
Wheat	6.4	3.7	11	10
Pulses	5.1	2.7	5	3
				$\frac{25.5}{1025.9} \frac{13.5}{504.2} \frac{15.3}{916.3} \frac{8.1}{450.1}$

$$\text{Laspeyre's index} = \frac{\sum p_1 q_1}{\sum p_0 q_0} \times 100 = \frac{504.2}{1025.9} \times 100 = 49.15$$

$$\text{Paasche's index} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{450.1}{916.3} \times 100 = 49.12$$

$$\text{Fisher's index} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 = \sqrt{49.15 \times 49.12} = 49.134$$

13.15

Commodity	1996		2000		p_0q_0	p_0q_1	p_1q_0	p_1q_1
	p_0	q_0	p_1	q_1				
A	2	74	3	82	148	165	222	246
B	5	125	4	140	625	700	500	560
C	7	40	6	33	280	231	240	198
					1053	1095	962	1004

$$\text{Marshall-Edgeworth index} = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100$$

$$= \frac{962 + 1004}{1053 + 1095} \times 100 = 91.53$$

$$\text{Fisher's Ideal index} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 = 91.523$$

13.16

Type of Crime	2000		2001		Weight, W	Crime Relative, R	RW
	W	R					
Robberies	13	8	6	$(8/13) \times 100$	369.24		
				= 61.54			
Car thefts	15	22	5	$(22/15) \times 100$	733.50		
				= 146.70			
Cycle thefts	249	185	4	$(185/249) \times 100$	297.16		
				= 74.29			
Pocket picking	328	259	1	$(259/328) \times 100$	78.96		
				= 78.96			
Thefts by servants	497	448	2	$(448/497) \times 100$	180.28		
				= 90.15			
		18			1659.14		

$$\text{Crime index} = \frac{\Sigma RW}{\Sigma W} = \frac{1659.14}{18} = 92.17$$

13.17

Commodity	Quantity		Price		
	1985 1993		1993		
	q_0	q_1	p_1	$q_1 p_1$	$q_0 p_1$
A	100	150	$900/150 = 6$	900	600
B	80	100	$500/100 = 5$	500	400
C	60	72	$360/72 = 5$	360	300
D	30	33	$297/33 = 9$	297	270
				2057	1570

$$\text{Paasche's quantity index} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100 = \frac{2057}{1570} \times 100$$

$$= 131.02$$

13.18

Commodity	Quantity		1995	Price	Percent Relatives	Base Value	Weighted Relatives
	1995	1999					
	q_0	q_1					
	(1)	(2)					
Wheat	29	24	3.80	83	$\frac{q_1}{q_0} \times 100$	110.20	9,146.60
Corn	3	2.5	2.91	83		8.73	724.59
Soyabeans	12	14	6.50	117	$q_0 p_0$	78.00	9,126.00
						196.93	18,997.19

Weighted average of relative quantity index

$$= \frac{\sum \{(q_1/q_0) \times 100\} (q_0 p_0)}{\sum q_0 p_0} = \frac{18,997.19}{196.93} = 96.$$

13.10 TESTS OF ADEQUACY OF INDEXES

So far we have discussed several methods to construct unweighted and weighted index numbers. However, the problem still remains of selecting an appropriate method for the construction of an index number in a given situation. The following tests have been suggested to select the adequacy of an index number:

- Time reversal test
- Factor reversal test
- Circular test

13.10.1 Time Reversal Test

The time reversal test is used to test whether a given method will work both backwards and forwards with respect to time. The test is that the formula for calculating the index number should be such that it will give the same ratio between one point of comparison and another no matter which of the two is taken as base. In other words, a price or quantity index for a given period with respect to the preceding period is equal to the reciprocal of the price or quantity index when periods are interchanged. For example, if P_{01} is a price index in the current year '1' with base of preceding year '0' and P_{10} is a price index for the base year '0' based on the current year '1', then the following relation should be satisfied:

$$P_{01} = \frac{1}{P_{01}} \quad \text{or} \quad P_{01} \times P_{10} = 1 \quad \text{and} \quad Q_{01} \times Q_{10} = 1$$

This test is not satisfied by the Laspeyre's Index and the Paasche's Index. The methods which satisfy the time reversal test are:

- Fisher's ideal index method
- Simple geometric mean of price relatives
- Aggregates with fixed weights (Kelly's formula)
- Marshall-Edgeworth's method
- Weighted geometric mean of price relatives when fixed weights are used
- Walsch's formula

For example, let us see how Fisher's ideal index formula satisfies the time reversal test.

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}}$$

For calculating P_{10} the time is interchanged so that p_0 becomes p_1 and p_1 becomes p_0 . Similarly q_0 becomes q_1 and q_1 becomes q_0 , we get

$$P_{10} = \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}} = 1$$

Since $P_{01} \times P_{10} = 1$, Fisher's ideal index satisfies the test.

13.10.2 Factor Reversal Test

According to Fisher, '*Just as each formula should permit the interchange of two items without giving inconsistent result so it ought to permit interchanging the prices and quantities without giving inconsistent result, i.e. the two results multiplied together should give the true value ratio.*' In other words, the change in the price when multiplied by the change in quantity should represent the total change in value. Thus, if the price of a commodity has doubled during a certain period and in this period the quantity has trebled, then the total change in the value should be six times the former level. That is, if p_1 and p_0 represent the prices and q_1 and q_0 the quantities in the current and the base periods respectively, then the price index for period '1' with base year '0' and the quantity index for period '1' with base year '0' is given by

$$P_{01} \times Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

The factor reversal test is satisfied only by the Fisher's ideal price index as shown below:

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}}$$

Changing p to q and q to p , we get the quantity index:

$$Q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}}$$

Multiplying P_{10} and Q_{01} , we get

$$\sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} = \sqrt{\frac{\sum p_1 q_1}{\sum p_0 q_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_0}} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

This result implies that the Fisher's index formula could be used for constructing both price and quantity indexes.

13.10.3 Circular Test

This test is concerned with the measurement of price change over a period of years when the shifting of base is desirable in a circular fashion. It may therefore be considered as an extension of time reversal test. For example, if an index is constructed for the year 2000 with the base of 1999 and another index for 1999 with the base of 1998, then it should be possible for us to directly get an index for the year 2000 with the base of 1998. If the index calculated directly does not give an inconsistent value, the circular test is said to be satisfied. If P_{ab} is price index for period 'b' with base period 'a'; P_{bc} is the price index for period 'c' with base period 'b' and P_{ca} is the price index for period 'a' with base period 'c', then an index is said to satisfy the circular test provided

$$P_{ab} \times P_{bc} \times P_{ca} = 1$$

This test is not satisfied by most of the common formulae used in the construction of indexes. Even Fisher's ideal formula does not satisfy this test. This test is satisfied by simple aggregative index, simple geometric mean of price relatives and weighted aggregate (with weight) index.

Example 13.14: Construct Fisher's price index using following data and show how it satisfies the time and factor reversal tests.

Commodity	2002		2003	
	Quantity	Price	Quantity	Price
A	20	12	30	14
B	13	14	15	20
C	12	10	20	15
D	8	6	10	4
E	5	8	5	6

Solution: Table 13.17 presents all the necessary information for constructing the Fisher's ideal index number.

Table 13.17: Calculation of Fisher's Ideal Index

Commodity	2002		2003		$p_1 q_0$	$p_0 q_0$	$q_1 p_1$	$p_0 q_1$
	q_0	p_0	q_1	p_1				
A	20	12	30	14	280	240	420	360
B	13	14	15	20	260	182	300	210
C	12	10	20	15	180	120	300	200
D	8	6	10	4	32	48	40	60
E	5	8	5	6	30	40	30	40
					782	630	1090	870

Fisher's ideal price index

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} = \sqrt{\frac{782}{630} \times \frac{1090}{870}} = 1.2471$$

Time Reversal Test: This test is satisfied when $P_{01} \times P_{10} = 1$

$$P_{10} = \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}} = \sqrt{\frac{870}{1090} \times \frac{630}{782}} = 0.8019$$

$$P_{01} \times P_{10} = \sqrt{\frac{782}{630} \times \frac{1090}{870} \times \frac{870}{1090} \times \frac{630}{782}} = \sqrt{1.2471 \times 0.8019} = 1$$

Hence, time reversal test is satisfied.

Factor Reversal Test: This test is satisfied when $P_{01} \times Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \text{ and } Q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}}$$

Thus, $P_{01} \times Q_{01} = \sqrt{\frac{782}{630} \times \frac{1090}{870} \times \frac{870}{630} \times \frac{1090}{782}} = \frac{1090}{630}$ which is the value of $\frac{\sum p_1 q_1}{\sum p_0 q_0}$

Hence, factor reversal test is also satisfied.

Example 13.15: Calculate Fisher's Ideal index from the data given below and show that it satisfies the time reversal test.

Commodity	2000		2001	
	Price	Quantity	Price	Quantity
A	10	49	12	50
B	12	25	15	20
C	18	10	20	12
D	20	5	40	2

Solution: Table 13.18 presents information necessary for Fisher's method to calculate the index.

Table 13.18: Calculation of Fisher's Ideal Index

Commodity	2000		2001		$p_1 q_0$	$p_0 q_0$	$p_1 q_1$	$p_0 q_1$
	p_0	q_0	p_1	q_1				
A	10	49	12	50	588	490	600	500
B	12	25	15	20	375	300	300	240
C	18	10	20	12	200	180	240	216
D	20	5	40	2	200	100	80	40
					1363	1070	1220	996

$$\text{Fisher's index } P_{01} = \sqrt{\frac{\sum p_1 q_0 \times \sum p_1 q_1}{\sum p_0 q_0 \times \sum p_0 q_1}} \times 100 = \sqrt{\frac{1363 \times 1220}{996 \times 1070}} \times 100 = 124.9$$

Time Reversal Test: This test is satisfied when $P_{01} \times P_{10} = 1$

$$P_{10} = \sqrt{\frac{\sum p_0 q_1 \times \sum p_0 q_0}{\sum p_1 q_1 \times \sum p_1 q_0}}$$

$$\begin{aligned} \text{Thus } P_{01} \times P_{10} &= \sqrt{\frac{\sum p_1 q_0 \times \sum p_1 q_1 \times \sum p_0 q_1 \times \sum p_0 q_0}{\sum p_0 q_0 \times \sum p_0 q_1 \times \sum p_1 q_1 \times \sum p_1 q_0}} \\ &= \sqrt{\frac{1363 \times 1220}{996 \times 1070} \times \frac{996}{1220} \times \frac{1070}{1363}} = \sqrt{1} = 1 \end{aligned}$$

Hence, the time reversal test is satisfied.

13.11 CHAIN INDEXES

The various formulae discussed so far assumed that the base period is any fixed period. The base period is the immediately preceding year of the current year. Moreover, the index of a given period on a given fixed base was not affected by changes in the relevant values of any other year. But in the chain base method, the data of each period is related with that of the immediately preceding period and not with any fixed period. This means that for the index of 2000 the base would be 1999 and for the index of 1999 the base would be 1998 and similarly of the index of 1998 the base would be 1997. Such index numbers are very useful in comparing current period data with the preceding period's data. Fixed base index in such a case does not give an appropriate comparison, because all prices are based on the fixed base period which may be far away for the current period and the preceding period.

For constructing an index by the chain base method, a series of indexes are computed for each period with preceding period as the base. These indexes are known as *link index* or *link relatives*. The steps of calculating link relatives are summarized below:

- (i) Express the data of a particular period as a percentage of the preceding period's data. This is called the *link relative*.
- (ii) These link relatives can be chained together. This is done by multiplying the link relative of the current year by the *chain index* of the previous year and dividing the product by 100. Thus

$$\text{Chain index for current year} = \frac{\text{Link relative of current year} \times \text{Chain index of previous year}}{100}$$

The chain index is useful for long-term comparison whereas link relatives are used for a comparison with the immediately preceding period. The fixed base indexes compiled from the original data and the chain indexes compiled from link relatives give the same value of index provided there is only one commodity whose indexes are being constructed.

Remarks Chain relatives differ from fixed base relatives in computation. Chain relatives are computed from link relatives whereas fixed base relatives are computed directly from original data.

$$\text{Link relative} = \frac{\text{Price relative for the current period}}{\text{Price relative for the previous period}} \times 100$$

$$\text{Price relative} = \frac{\text{Current period's link relative} \times \text{Previous period's price relatives}}{100}$$

Multiplying the link relatives $P_{01}, P_{12}, P_{23}, \dots, P_{(n-1)n}$ successively is known as the chaining process that gives link relatives with a common base:

$$P_{01} = \text{First link}$$

$$P_{02} = P_{01} \times P_{12}$$

$$P_{03} = (P_{01} \times P_{12}) \times P_{23} = P_{02} \times P_{23}$$

.

.

$$P_{0n} = P_{0(n-1)} \times P_{(n-1)n}$$

Conversion of chain base index (CBI) to fixed base index (FBI)

$$\text{Current period FBI} = \frac{\text{Current period CBI} \times \text{Previous period FBI}}{100}$$

Example 13.16: Construct an index by the chain base method based on the following data of the wholesale prices of a certain commodity.

Year	Price	Year	Price
1994	37	2000	48
1995	39	2001	49
1996	43	2002	54
1997	48	2003	56
1998	48	2004	87
1999	52		

Solution: Computation of the chain base index number is shown in Table 13.19.

Table 13.19: Chain Base Indexes

Year	Price	Link Relative	Chain Base Index Numbers (Base Year 1985 = 100)
1994	37	100.00	100
1995	39	$(39/37) \times 100 = 105.41$	$(105.41/100) \times 100 = 105.41$
1996	43	$(43/39) \times 100 = 110.26$	$(110.76/100) \times 105.41 = 116.23$
1997	48	$(48/43) \times 100 = 111.63$	$(111.63/100) \times 116.23 = 129.75$
1998	48	$(48/48) \times 100 = 100.00$	$(100/100) \times 129.75 = 129.75$
1999	52	$(52/48) \times 100 = 108.33$	$(108.33/100) \times 129.75 = 140.56$
2000	48	$(48/48) \times 100 = 100.00$	$(100/100) \times 140.56 = 140.56$
2001	49	$(49/48) \times 100 = 102.08$	$(102.08/100) \times 140.56 = 143.48$
2002	54	$(54/49) \times 100 = 110.20$	$(110.20/100) \times 158.11 = 158.11$
2003	56	$(56/54) \times 100 = 103.70$	$(103.70/100) \times 158.11 = 163.96$
2004	57	$(57/56) \times 100 = 101.79$	$(110.79/100) \times 163.96 = 166.90$

Example 13.17: Prepare fixed base index numbers from the chain base index numbers given below:

Year	:	1996	1997	1998	1999	2000	2001
Chain Index :		94	104	104	93	103	102

Solution: Computation of fixed base indexes is shown in Table 13.20 using the following formula:

$$\text{Fixed base index (FBI)} = \frac{\text{Current period CBI} \times \text{Previous period FBI}}{100}$$

Table 13.20: Fixed Base Index Numbers

Year	Chain Index	Fixed Base Index
1996	94	94
1997	104	$\frac{104 \times 94}{100} = 97.76$
1998	104	$\frac{104 \times 97.76}{100} = 101.67$
1999	93	$\frac{93 \times 101.67}{100} = 94.55$
2000	103	$\frac{103 \times 94.55}{100} = 97.39$
2001	102	$\frac{102 \times 97.39}{100} = 99.34$

Example 13.18: Calculate the chain base index number and fixed base index number from the following data:

Commodity	1998	1999	2000	2001	2002
A	4	6	8	10	12
B	16	20	24	30	36
C	8	10	16	20	24

Solution: Computation of the chain index number and fixed base index number is shown in Tables 13.21 and 13.22.

Table 13.21: Chain Base Index Number

Commodity	Link Relatives Based on Preceding Year				
	1998	1999	2000	2001	2002
A	100	$\frac{6}{4} \times 100 = 150$	$\frac{8}{6} \times 100 = 133.33$	$\frac{10}{8} \times 100 = 125$	$\frac{12}{10} \times 100 = 120$
B	100	$\frac{20}{16} \times 100 = 125$	$\frac{24}{20} \times 100 = 120$	$\frac{30}{24} \times 100 = 125$	$\frac{36}{30} \times 100 = 120$
C	100	$\frac{10}{8} \times 100 = 125$	$\frac{16}{10} \times 100 = 160$	$\frac{20}{16} \times 100 = 125$	$\frac{24}{20} \times 100 = 120$
Total link relatives	300	400	413.33	375	360
Average link relatives	100	133.33	137.78	125	120
Chain base index	100	$\frac{133.33 \times 100}{100} = 133.33$	$\frac{137.78 \times 133.33}{100} = 183.70$	$\frac{125 \times 183.70}{100} = 229.63$	$\frac{120 \times 229.63}{100} = 275.56$

Table 13.22: Fixed Base Index Number

Commodity	Price Relatives (Base 1998 = 100)				
	1998	1999	2000	2001	2002
A	100	$\frac{6}{4} \times 100 = 150$	$\frac{8}{4} \times 100 = 200$	$\frac{10}{4} \times 100 = 250$	$\frac{12}{4} \times 100 = 300$
B	100	$\frac{20}{16} \times 100 = 125$	$\frac{24}{16} \times 100 = 150$	$\frac{30}{16} \times 100 = 187.5$	$\frac{36}{16} \times 100 = 225$
C	100	$\frac{10}{8} \times 100 = 125$	$\frac{16}{8} \times 100 = 200$	$\frac{20}{8} \times 100 = 250$	$\frac{24}{8} \times 100 = 300$
Total	300	400	550	687.5	825
Average (Fixed base index number)	100	133.33	183.33	229.17	275

Advantages and Disadvantages of Chain Base Indexes

Advantages: The following are a few advantages of chain base index

- (i) The chain base indexes enable us to make comparisons with the *previous and not any distant past period*. Thus these index are very useful in the analysis of business data.
- (ii) The chain base method permits us to introduce new commodities and delete the existing ones which are obsolete without any recalculations of the entire series.
- (iii) The index numbers calculated by the chain base method are relatively free from cyclical and seasonal variations.

Disadvantage: The main disadvantage of the chain base index is that it is not useful for long term comparisons of chained percentages in a time series. The process of chaining link relatives is computationally difficult.

13.12 APPLICATIONS OF INDEX NUMBERS

13.12.1 Changing the Base of an Index

The reasons for changing the base of an index number and constructing a new index number based on a new base year are:

1. The base year is either too old or distant from the current year. Consequently it becomes useless for meaningful comparison of prices or commodities in the two different periods. For example, if prices of 2002 are compared with prices of 1970, then such comparison is useless. It is desirable that the base period should be a period of stability and should be such around which the prices of the current year fluctuate.
2. A comparison is to be made with a series of indexes with a different base. The method of base shifting requires considering the new base year as 100 and then expressing all the indexes as percentages of the index of the new base period selected. The formula for shifting the base is:

$$\begin{aligned}\text{New index of any year} &= \frac{\text{Old index of the period}}{\text{Old index of the new base}} \times 100 \\ &= \left(\frac{100}{\text{Index of the new base period}} \right) \times \text{Old index of the period}\end{aligned}$$

In other words, a new index number is obtained by multiplying the old index with a common factor: $\{(100/\text{Index number of new base})\}$. The procedure gives correct results only when the index number satisfies the circular test.

Example 13.19: The following are the indexes of prices based on 1993 prices. Shift the base from 1993 to 1998.

Year:	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Index:	100	110	120	200	320	400	410	400	380	370	350	366

Solution: The method of shifting the base period from 1993 to 1998 is demonstrated in Table 13.23.

Table 13.23: Shifting of Base Period

	Year	Old Index (1995 = 100)	New Index (1998 = 100)
Old base year	1993	100	$(100/400) \times 100 = 25$
	1994	110	$(110/400) \times 100 = 27.5$
	1995	120	$(120/400) \times 100 = 30$
	1996	200	$(200/400) \times 100 = 50$
	1997	320	$(320/400) \times 100 = 80$
New base year	1998	400	100
	1999	410	$(410/400) \times 100 = 102.5$
	2000	400	$(400/400) \times 100 = 100$
	2001	380	$(380/400) \times 100 = 95$
	2002	370	$(370/400) \times 100 = 92.5$
	2003	350	$(350/400) \times 100 = 87.5$
	2004	366	$(366/400) \times 100 = 91.5$

Example 13.20: An index is at 100 in 1995. It rises 4 per cent in 1996, falls 6 per cent in 1997, falls 4 per cent in 1998, and rises 3 per cent in 1999. Calculate the index number for 5 years with 1997 as the base.

Solution: Calculations of index at the old base year 1995 and then construction the new indexes with 1997 as the base year is shown in Table 13.24.

Table 13.24: Price Relatives

Year	<i>Old Index (1995 = 100)</i>	<i>New Index (1997 = 100)</i>
1995	100	$\frac{100}{97.76} \times 100 = 102.32$
1996	$100 + 4 = 104$	$\frac{100}{97.76} \times 104 = 106.40$
1997	$\frac{94}{100} \times 104 = 97.76$	100.00
1998	$\frac{96}{100} \times 97.76 = 93.85$	$\frac{100}{97.76} \times 93.85 = 96.00$
1999	$\frac{103}{100} \times 93.85 = 96.66$	$\frac{100}{97.76} \times 96.66 = 98.88$

13.12.2 Combining Two or More Overlapping Indexes

Two or more overlapping indexes are combined to obtain a single index on a common base. It is also called *splicing* or *coupling* of indexes. The need to combine arises for maintaining continuity in comparison because sometimes an index is discontinued as soon as its base becomes too old. A new series of indexes is constructed with a recent period as base by the same technique as used in base shifting. Splicing of indexes can be done only if the indexes are constructed with the same items, and have an overlapping year. Suppose we have an index number with a base of 1991 and another index (using the same items as the first one) with a base of 2001. If both indexes are continuing, then we can splice the first index to the second one and have a common index with base year 2001 for all years. We can also splice the index of 2001 with the index of 1991 and have a common index with base year 1991.

Two indexes with different bases are spliced (coupled) into a continuous series of indexes with a common base period using the following two approaches.

Forward splicing approach This approach is used for splicing an old series of indexes to make it continuous with the new series of indexes. The formula used is:

$$\text{Required inde} = \frac{(\text{Old index with existing base}) \times (\text{An index to be spliced})}{100}$$

Backward splicing approach This approach is used for splicing a new series of indexes to make it continuous with an old series. The formula used is:

$$\text{Required index} = \frac{100}{\text{Old index with existing base}} \times \text{An index to be spliced}$$

Example 13.21: Given below are two price indexes. Splice them on the base 1999 = 100. By what percentage did the price of the commodity rise between 1995 and 2000?

Year	<i>Old Price Index (1990 = 100)</i>	<i>New Price Index (1999 = 100)</i>
1995	141.5	—
1996	163.7	—
1997	158.2	—
1998	156.8	99.8
1999	157.1	100.0
2000	—	102.3

Solution: The old index is to be spliced to the new one or the new index has to go back to the data of the old index. This is backward splicing. Here the common factor for splicing would be 100 divided by the old index on the new base year as shown in Table 13.25.

Table 13.25: Splicing Old Index to the New Index

Year	Old Price Index Base (1990 = 100)	New Price Index Base (1999 = 100)
1995	141.5	$\frac{100}{157.1} \times 141.5 = 90.06$
1996	163.7	$\frac{100}{157.1} \times 163.7 = 104.19$
1997	158.2	$\frac{100}{157.1} \times 158.2 = 100.69$
1998	156.8	$\frac{100}{157.1} \times 156.8 = 99.00$
1999	157.1	— 100
2000	—	— 102.3

The price rise between 1995 and 2000 in terms of percentage is given by

$$\frac{102.3 - 90.06}{90.06} \times 100 = \frac{12.24}{90.06} \times 100 = 13.59 \text{ per cent}$$

Example 13.22: A price index series was started in 1987 as base. By 1991 it rose by 25 per cent. The link relative for 1992 was 95. In this year a new series was started. This new series rose by 15 points by the next year. But during the next four years the rise was not rapid. During 1997 the price level was only 5 per cent higher than 1995, and in 1995 they were 8 per cent higher than 1993. Splice the two series and calculate the indexes numbers for the various years shifting the base to 1992.

Solution: The calculations for splicing of indexes and base shifting are shown in Tables 13.26 and 13.27, respectively.

Table 13.26: Splicing of the Index Numbers

Year	Index No. (1987 = 100)	Index No. (1992 = 100)	Old Index Spliced to New One (1992 = 100)
1987	100	—	$\frac{100}{118.75} \times 100 = 84.21$
1991	125	—	$\frac{100}{118.75} \times 125 = 105.26$
1992	$\left(\frac{95}{100} \times 125 \right) = 118.75$	100	100
1993	—	115	115
1995	—	$\left(\frac{115 \times 108}{100} \right) = 124.2$	124.2
1997	—	$\left(\frac{124.2 \times 105}{100} \right) = 130.41$	130.41

Table 13.27: Base Shifting to 1993

Year	Index Number
1987	$\frac{100}{115} \times 84.21 = 73.23$
1991	$\frac{100}{115} \times 105.26 = 91.53$
1992	$\frac{100}{115} \times 100 = 86.96$
1993	$\frac{100}{115} \times 115 = 100.00$
1995	$\frac{100}{115} \times 124.2 = 108.00$
1997	$\frac{100}{115} \times 130.41 = 113.40$

13.12.3 Correction (Adjustment) of Value of an Item

Correcting or adjusting the rupee value of an item at a time period (also called *deflating*) is achieved by dividing it by the appropriate price index of the same time period after considering changes in price levels. When prices rise, the purchasing power of money declines. If the money incomes of people remain constant and prices are doubled then the purchasing power of money is reduced to half. For example, in 1995 a person was getting Rs 1000 per month and he could purchase 10 units of an item at the rate of Rs 100 per unit. However, in the year 2000 the price of the item has increased to Rs 125, therefore the person could buy only $100/125 = 8$ units of the item.

When prices increase, the money wages are deflated by the price index to get the figure of real wages. The real wages enable us to see whether a wage earner is better-off or worse-off as a result of a price change. Thus real wage or income is determined by using the formula:

$$\text{Real wage or income} = \frac{\text{Money wage}}{\text{Price index}} \times 100$$

Here the price index should be the consumer price index as it would reflect the change in purchasing power of the wage earner. Thus

$$\begin{aligned}\text{Real wage index} &= \frac{\text{Real wage of current year}}{\text{Real wage of base year}} \times 100 \\ &= \frac{\text{Index of money wage}}{\text{Consumer price index}} \times 100\end{aligned}$$

Example 13.23: The employees of an organization have presented the following data in support of their contention that they are entitled to a wage adjustment. The following data represent the average monthly take-home salary of the employees:

Year	1998	1999	2000	2001
Pay	10,420	10,432	10,960	11,300
Index	126.8	129.5	136.2	141.2

- (a) Compute the real wages based on the take-home pay and the price indexes given.
- (b) Compute the amount of pay needed in 2001 to provide buying power equal to that enjoyed in 1998.

Solution: (a) The calculations for real wages based on take-home salary and price indexes are shown in Table 13.28.

Table 13.28: Computation of Real Wages

Year (1)	Salary (in Rs) (2)	Price Index (3)	Real Wages (4) = $\{(2) \div (3)\} \times 100$
1998	10,420	126.8	$\frac{10,420}{126.8} \times 100 = 8217.66$
1999	10,432	129.5	$\frac{10,432}{129.5} \times 100 = 8055.59$
2000	10,960	136.2	$\frac{10,960}{136.2} \times 100 = 8046.98$
2001	11,300	141.2	$\frac{11,300}{141.2} \times 100 = 8002.83$

- (b) In order that the employees have the same buying power in 2001 as they had in 1998, the pay in 2001 should be

$$\frac{10,420}{126.8} \times 141.2 = \text{Rs } 11,603.34$$

Example 13.24: Given below are the average wages in rupees per hour of unskilled workers of a factory during the period 1995–2000. Also shown is the consumer price index of these years (taking 1995 as base year with price index = 100). Determine the real wages of the workers during 1995–2000 compared with their wages in 1995.

Year	1995	1996	1997	1998	1999	2000
Consumer price index :	100	120.2	121.7	125.9	129.2	140
Average wage (Rs/hr) :	11.9	19.4	21.3	22.8	24.5	31.0

How much is the worth of one rupee of 1995 in subsequent years?

Solution: Calculations for real wages of workers and worth of one rupee of 1995 in subsequent years are shown in Table 13.29.

Table 13.29: Worth of One Rupee of 1995

Year	Consumer Price Index (CPI) 1995 = Base Period	Average Wage (Rs/hr)	Real Wage of Workers (Rs/hr)	Worth of Re 1 in 1995
1995	100	11.9	$\frac{11.9}{100} \times 100 = 11.90$	$\frac{100}{100} = 1.00$
1996	120.2	19.4	$\frac{19.4}{120.2} \times 100 = 16.13$	$\frac{100}{120.2} = 0.83$
1997	121.7	21.3	$\frac{21.3}{121.7} \times 100 = 17.50$	$\frac{100}{121.7} = 0.83$
1998	125.9	22.8	$\frac{22.8}{125.9} \times 100 = 18.10$	$\frac{100}{125.9} = 0.79$
1999	129.3	24.5	$\frac{24.5}{129.3} \times 100 = 18.94$	$\frac{100}{129.3} = 0.77$
2000	140.0	31.0	$\frac{31.0}{140.0} \times 100 = 22.14$	$\frac{100}{140} = 0.71$

Self-Practice Problems 13C

- 13.19** Calculate Fisher's Ideal index from the data given below and show that it satisfies the time reversal and factor reversal tests.

Commodity	Base Year		Current Year	
	Quantity	Price	Quantity	Price
A	12	10	15	12
B	15	7	20	5
C	24	5	20	9
D	5	16	5	14

- 13.20** Splice the following two index number series, continuing series A forward and series B backward.

Year	1998	1999	2000	2001	2002	2003
Series A :	100	120	150	—	—	—
Series B :	—	—	100	110	120	150

- 13.21** Calculate the chain base index number chained to 1994 from the average price of following three commodities:

Commodity	1999	2000	2001	2002	2003
Wheat	4	6	8	10	12
Rice	16	20	24	30	36
Sugar	8	10	16	20	24

- 13.22** The following table gives the annual income of a clerk and the general index number of price during 1994–98. Prepare the index number to show the changes in the real income of the teacher.

Year	Income (Rs)	Price Index No.	Year	Income (Rs)	Price Index No.
			1994	1995	1996
1994	36,000	100	1999	64,000	290
1995	42,000	104	2000	68,000	300
1996	50,000	115	2001	72,000	320
1997	55,000	160	2002	75,000	330
1998	60,000	280	—	—	—

- 13.23** Calculate Fisher's Ideal index number from the given data. Does it satisfy the time reversal and factor reversal tests?

Commodity	Price	Quantity	Price	Quantity
A	6	50	10	56
B	2	100	2	120
C	4	60	6	60
D	10	30	12	24
E	8	40	12	36

[Bangalore Univ., BCom, 1998]

- 13.24** From the following average price of groups of commodities given in rupees per unit, find the chain base index number with 1994 as the base year:

Group	1994	1995	1996	1997	1998
I	2	3	4	5	6
II	8	10	12	15	18
III	4	5	18	10	12

[Agra Univ., BCom, 1998]

- 13.25** Given the following data:

Year	Weekly Take-home Pay (Wages)	Consumer Price Index
1998	109.50	112.8
1999	112.20	118.2
2000	116.40	127.4
2001	125.08	138.2
2002	135.40	143.5
2003	138.10	149.8

- (a) What was the real average weekly wage for each year?
- (b) In which year did the employees have the greatest buying power?
- (c) What percentage increase in the weekly wages for the year 2003 is required (if any) to provide the

same buying power that the employees enjoyed in the year in which they had the highest real wages?

- 13.26** Using the following data construct Fisher's Ideal index and show that it satisfies the factor reversal and time reversal tests:

Commodity	Price (in Rs/Unit)		Number of Units	
	Base Year	Current Year	Base Year	Current Year
A	6	8	10	12
B	10	10	5	8
C	5	7	8	10
D	15	20	12	15
E	20	25	15	10

- 13.27** From the data given below, calculate Fisher's Ideal index and show that it satisfies the time reversal and factor reversal test:

Commodity	1998		1999	
	Price	Quantity	Price	Quantity
A	12	20	14	30
B	14	13	20	15
C	10	12	15	20
D	6	8	4	10
E	8	5	6	5

[Sukhadia Univ., MBA, 1999]

- 13.28** Calculate the index number by using Paasche's method, and Fisher's method.

Commodity	p_1	q_1	p_0	q_0
A	5	14	3	8
B	8	18	6	25
C	3	25	1	40
D	15	36	12	48
E	9	14	7	18
F	7	13	5	19

[MD Univ., MBA, 1996]

Hints and Answers

- 13.19**

Commodity	q_0	p_0	q_1	p_1	$p_1 q_0$	$p_0 q_0$	$p_1 q_1$	$p_0 q_1$
A	12	10	15	12	144	120	180	150
B	15	7	20	5	75	105	100	140
C	24	5	20	9	216	120	180	100
D	5	16	5	14	70	80	70	80
					505	425	530	470

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} = \sqrt{\frac{505}{425} \times \frac{530}{470}} = 1.1576$$

$$P_{10} = \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}} = \sqrt{\frac{470}{530} \times \frac{425}{505}} = 0.8638$$

Time Reversal Test

$$P_{01} \times P_{10} = \sqrt{\frac{505}{425} \times \frac{530}{670} \times \frac{470}{530} \times \frac{425}{505}} = \sqrt{1} = 1$$

Factor Reversal Test

$$P_{01} \times Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

$$Q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} = \sqrt{\frac{470}{525} \times \frac{530}{505}} = 0.9693$$

$$P_{01} \times Q_{01} = \sqrt{\frac{505}{425} \times \frac{530}{470} \times \frac{470}{425} \times \frac{530}{505}} = \sqrt{\frac{530}{425} \times \frac{530}{425}}$$

$$= \frac{503}{425} \text{ which is equal to } \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

13.20

Year	Series A	Series B	Series B Spliced to A	Series A Spliced to B
1998	100	—	—	$\left(\frac{100}{150}\right) \times 100 = 66.66$
1999	120	—	—	$\left(\frac{100}{150}\right) \times 120 = 80.00$
2000	150	100	$\left(\frac{150}{100}\right) \times 100 = 150$	$\left(\frac{100}{150}\right) \times 150 = 100.00$
2001	—	110	$\left(\frac{150}{100}\right) \times 110 = 165$	
2002	—	120	$\left(\frac{150}{100}\right) \times 120 = 180$	
2003	—	150	$\left(\frac{150}{100}\right) \times 150 = 225$	

13.22

Year	Income (Rs)	Price Index	Real Income	Real Income Index
1994	360	100	$(360/100) \times 100 = 360.00$	100.00
1995	420	104	$(420/104) \times 100 = 403.85$	112.18
1996	500	115	$(500/115) \times 100 = 434.78$	120.77
1997	550	160	$(550/160) \times 100 = 343.75$	95.49
1998	600	280	$(600/280) \times 100 = 214.29$	59.52
1999	640	290	$(640/290) \times 100 = 220.69$	61.30
2000	680	300	$(680/300) \times 100 = 226.67$	62.96
2001	720	320	$(720/320) \times 100 = 225.00$	62.52
2002	750	330	$(750/330) \times 100 = 227.27$	63.13

13.21

Commodity	Relatives Based on the Preceding Year				
	1999	2000	2001	2002	2003
Wheat	100	150	133.33	125	120
Rice	100	125	120.00	125	120
Sugar	100	125	160.00	125	120
Total	300	400	413.33	375	360
Average of link relatives	100	133.33	137.78	125	120
Chain index	100	$\frac{133.33 \times 100}{100}$	$\frac{137.78 \times 133.33}{100}$	$\frac{125 \times 183.70}{100}$	$\frac{120 \times 229.63}{100}$
(1999 = 100)		= 133.33	= 183.70	= 229.63	= 275.55

13.24

Group	1994		1995		1996		1997		1998	
	Price	Link Relative	Price	Link Relative	Price	Link Relative	Price	Link Relative	Price	Link Relative
I	2	100	3	150	4	133.3	5	125	6	120
II	8	100	10	125	12	120.0	15	125	18	120
III	8	100	5	125	8	160.0	10	125	12	120
Total	300		400		413.3		375		360	
Average of link relatives	100		133.33		137.77		125		120	
Chain index (1994 = 100)	100		$\frac{133.33}{100} \times 100$		$\frac{137.77}{100} \times 133.33$		$\frac{125}{100} \times 183.69$		$\frac{120}{100} \times 229.61$	
			= 133.33		= 183.69		= 229.61		= 275.53	

- 13.25** (a) Average weekly wage can be obtained by using the following formula:

$$\text{Real wage} = \frac{\text{Money wage}}{\text{Price index}} \times 100$$

Year	Weekly Take-home Pay (Rs)	Consumer Price Index	Real wages
1998	109.50	112.8	$\frac{109.5}{112.8} \times 100 = 97.07$
1999	112.20	118.2	$\frac{112.2}{118.2} \times 100 = 94.92$
2000	116.40	127.4	$\frac{116.4}{127.4} \times 100 = 91.37$
2001	125.08	138.2	$\frac{125.08}{138.2} \times 100 = 90.51$
2002	135.40	143.5	$\frac{135.4}{143.5} \times 100 = 94.36$
2003	138.10	149.8	$\frac{138.10}{149.8} \times 100 = 92.19$

- (b) Since real wage was maximum in the year 1998, the employees had the greatest buying power in that year.
(c) The percentage increase in the weekly wages for the year 2003 required to provide the same buying power that the employees had in 1998:
Absolute difference = $97.07 - 92.19 = 6.88$.

13.26

Comm	p_0	p_1	q_0	q_1	$p_1 q_0$	$p_0 q_0$	$p_1 q_1$	$p_0 q_1$
A	6	8	10	12	80	60	96	72
B	10	10	5	8	50	50	80	80
C	5	7	8	10	56	40	70	60
D	15	20	12	15	240	180	300	225
E	20	25	15	10	375	300	250	200
					801	630	796	627

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}};$$

$$P_{10} = \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}}$$

Time Reversal Test: $P_{01} \times P_{10} = 1$

$$P_{01} \times P_{10} = \sqrt{\frac{801}{630} \times \frac{796}{627} \times \frac{627}{796} \times \frac{630}{801}} = \sqrt{1} = 1$$

$$\text{Factor Reversal Test: } P_{01} \times Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0} = \frac{796}{630}$$

$$Q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_0 p_1}{\sum q_1 p_1}}$$

$$= \sqrt{\frac{627}{630} \times \frac{796}{801}} P_{01} \times Q_{01}$$

$$= \sqrt{\frac{801}{630} \times \frac{796}{627} \times \frac{627}{630} \times \frac{796}{801}} = \frac{796}{630}$$

which is equal to $\frac{\sum p_1 q_1}{\sum p_0 q_0}$

13.27

Comm	1998		1999		$p_1 q_0$	$p_0 q_0$	$p_1 q_1$	$p_0 q_1$
	p_0	q_0	p_1	q_1				
A	12	20	14	30	280	240	420	360
B	14	13	20	15	260	182	300	210
C	10	12	15	20	180	120	300	200
D	6	8	4	10	32	48	40	60
E	8	5	6	5	30	40	30	40
					782	630	1090	870

Fisher's Ideal Index:

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

$$= \sqrt{\frac{782}{630} \times \frac{1090}{870}} \times 100 = 124.7$$

Time Reversal test:

$$P_{01} \times P_{10} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}}$$

$$= \sqrt{\frac{782}{630} \times \frac{1090}{870} \times \frac{870}{1040} \times \frac{630}{782}} = 1$$

13.28

Comm	p_1	q_1	p_0	q_0	$p_1 q_1$	$p_0 q_1$	$p_1 q_0$	$p_0 q_0$
A	5	14	3	8	70	42	40	24
B	8	18	6	25	144	108	200	150
C	3	25	1	40	75	25	120	40
D	15	36	12	48	540	432	720	576
E	9	14	7	18	126	98	162	126
F	7	13	5	19	91	65	133	95
					1046	770	1375	1011

(i) Paasche's Index,

$$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{1046}{770} \times 100 = 135.84$$

(ii) Fisher's Index,

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

$$= \sqrt{\frac{1375}{1011} \times \frac{1046}{770}} \times 100 = 135.92$$

13.13 CONSUMER PRICE INDEXES

The consumer price index, also known as the *cost of living index* or *retail price index*, is constructed to measure the amount of money which consumers of a particular class have to pay to get a basket of goods and services at a particular point of time in comparison to what they paid for the same in the base year.

The need for constructing consumer price indexes arises because the general indexes do not highlight the effects of rise or fall in prices of various commodities consumed by different classes of people on their cost of living. Moreover, different classes of people consume different types of commodities and even the same type of commodities are not consumed in the same proportion by different classes of people. To study the effect of rise or fall in prices of different types of commodities, the Cost of Living Index (CLI) are constructed separately for different classes of people.

The problem in constructing consumer price indexes arise because variations in prices of commodities have to be studied from the point of view of consumers living in different regions or places. Since retail prices in different places differ and the pattern of consumption is also not identical at different places, therefore people living in different regions, pay different prices to purchase various commodities. Moreover, the relative importance of various commodities to all people is not identical. Therefore we cannot construct one CLI for the whole country.

13.13.1 Uses of Consumer Price Index (CPI) Number

The importance of the CPI can be seen from the following:

- (i) The CPI are used to formulate economic policy, escalate income payments, and measure real earnings.
- (ii) The CPI are used to measure purchasing power of the consumer in rupees. The purchasing power of the rupee is the value of a rupee in a given year as compared to a base year. The formula for calculating the purchasing power of the rupee is:

$$\text{Purchasing power} = \frac{1}{\text{Consumer price index}} \times 100$$

- (iii) When a time series is concerned with such rupee values as retail sales amounts or wage rates, the price index is most frequently used to achieve deflation of such time-series. The process of deflating can be expressed in the form of a formula as:

$$\text{Real wage} = \frac{\text{Money value}}{\text{Consumer price index}} \times 100$$

- (iv) The CPI is used in wage negotiations and wage contracts. Automatic adjustment of wages or the dearness allowance component of the wages is done on the basis of the consumer price index.

13.13.2 Construction of a Consumer Price Index

The CPI is a weighted aggregate price index with fixed weights. The need for weighting arises because the relative importance of various commodities or items for different classes of people is not the same. The percentage of expenditure on different commodities by an average family constitutes the individual weights assigned to the corresponding price relatives, and the percentage expenditure on five well-accepted groups of commodities namely: (i) food, (ii) clothing, (iii) fuel and lighting, (iv) house rent, (v) miscellaneous.

The weight applied to each commodity in the market basket is derived from a usage survey of families throughout the country. The consumer price index or cost of living index numbers are constructed by the following two methods:

Aggregate expenditure method or weighted aggregate method

This method is similar to the Laspeyre's method of constructing a weighted index. To

apply this method, the quantities of various commodities consumed by a particular class of people are assigned weights on the basis of quantities consumed in the base year. Mathematically it is stated as:

$$\begin{aligned}\text{Consumer price index} &= \frac{\text{Total expenditure in current period}}{\text{Total expenditure in base period}} \times 100 \\ &= \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100\end{aligned}$$

where p_1 and p_0 = prices in the current period and base period, respectively

q_0 = quantities consumed in the base period

Family budget method or method of weighted average of price relatives

To apply this method the family budget of a large number of people, for whom the index is meant, are carefully studied. Then the aggregate expenditure of an average family on various commodities is estimated. These values constitute the weights. Mathematically, consumer price index is stated as:

$$\text{Consumer price index} = \frac{\sum PV}{\sum V} \times 100$$

when P = price relatives, $p_1/p_0 \times 100$

V = Value weight, $p_0 q_0$

Example 13.25: Owing to change in prices the consumer price index of the working class in a certain area rose in a month by one quarter of what it was prior to 225. The index of food became 252 from 198, that of clothing from 185 to 205, of fuel and lighting from 175 to 195, and that of miscellaneous from 138 to 212. The index of rent, however, remained unchanged at 150. It was known that the weight of clothing, rent and fuel, and lighting were the same. Find out the exact weight of all the groups.

[Delhi, Univ., MBA, 1997]

Solution: Suppose the weights of items included in the group are as follows:

- Food x
- Fuel and Lighting z
- Miscellaneous y
- Clothing z
- Rent z

Therefore, the weighted index in the beginning of the month would be:

	Index <i>I</i>	Weight <i>W</i>	<i>IW</i>
Food	198	x	$198x$
Clothing	185	z	$185z$
Fuel and Lighting	175	z	$175z$
Rent	150	z	$150z$
Miscellaneous	138	y	$138y$
		$x + y + 3z$	$198x + 138y + 510z$

$$\text{Index number} = \frac{198x + 138y + 510z}{x + y + 3z}$$

Similarly the weighted index at the end of the month would be:

	<i>I</i>	<i>W</i>	<i>IW</i>
Food	252	x	$252x$
Clothing	205	z	$205z$
Fuel and Lighting	195	z	$195z$
Rent	150	z	$150z$
Miscellaneous	212	y	$212y$
		$x + y + 3z$	$252x + 212y + 550z$

$$\text{Index number} = \frac{252x + 212y + 550z}{x + y + 3z}$$

The weighted index at the end of the month was 225 (given). This index is a rise from the first index by one quarter. Therefore, the index at the beginning was $(4/5)$ th of 225 = 180.

Hence the weighted index at the beginning of the month was

$$180 = \frac{198x + 138y + 510z}{x + y + 3z}$$

$$180 + 180y + 540z = 198x + 138y + 510z$$

$$18x - 42y - 30z = 0 \quad (\text{i})$$

Similarly the weighted index at the end of month was

$$225 = \frac{252x + 212y + 550z}{x + y + 3z}$$

$$225x + 225y + 675z = 252x + 212y + 550z$$

$$27z - 13y - 125z = 0 \quad (\text{ii})$$

Let the total weight be equal to 100. Hence

$$x + y + 3z = 100 \quad (\text{iii})$$

Multiplying Eqn. (iii) by 18 and subtracting from (i), we get

$$-60y - 84z = -1800 \text{ or } 60y + 84z = 1800 \quad (\text{iv})$$

Multiplying (iii) by 27, and subtracting from Eqn. (ii), we get

$$-40y - 206z = -2700 \text{ or } 40y + 206z = 2700$$

Multiplying Eqn. (iv) by 20, and Eqn. (v) by 30 and subtracting, we get

$$-4500z = -45000 \text{ or } z = 10$$

Substituting the value of z in Eqn. (iv), we have

$$60y + (84 \times 10) = 180 \text{ or } y = 10$$

Substituting the value of y and z in Eqn. (iii), we have

$$x + 16 + (3 \times 10) = 100 \text{ or } x = 54$$

Thus, the exact weights are:

- | | | | |
|---------------------|----|------------|----|
| • Food | 54 | • Clothing | 10 |
| • Fuel and Lighting | 10 | • Rent | 10 |
| • Miscellaneous | 16 | | |

Example 13.26: Calculate the index number using (a) Aggregate expenditure method, and (b) Family budget method for the year 2000 with 1995 as the base year from the following data:

Commodity	Quantity (in Units)	Price (in Rs/Unit)	
		1990	2000
A	100	8.00	12.00
B	25	6.00	7.50
C	10	5.00	5.25
D	20	48.00	52.00
E	25	15.00	16.50
F	30	9.00	27.00

Solution: Calculations of cost of living index are shown in Tables 13.30 and 13.31.

(a)

Table 13.30: Index Number by Aggregative Expenditure Method

Commodity	Price (Rs per unit) in		Quantity (in units) in 1990			
	1990			2000		
	p_0	p_1		q_0	$p_0 q_0$	
A	8.00	12.00	100	800.00	1200.00	
B	6.00	7.50	25	150.00	187.50	
C	5.00	5.25	10	50.00	52.50	
D	48.00	52.00	20	960.00	1040.00	
E	15.00	16.50	25	375.00	412.50	
F	9.00	27.00	30	270.00	810.00	
				2605.00	3702.50	

$$\text{Cost of living index} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{3702.5}{2605} \times 100 = 142.13$$

(b)

Table 13.31: Index Number by Family Budget Method

Commodity	Price (Rs per unit) in		Quantity (in units) in 1990	Price Relatives $P (P_1/p_0) \times 100$	Weights		
	1990				$W = p_0 q_0$	PW	
	p_0	p_1	q_0				
A	8.00	12.00	100	150.00	800	1,20,000	
B	6.00	7.50	25	125.00	150	18,750	
C	5.00	5.25	10	105.00	50	5,250	
D	48.00	52.00	20	108.33	960	1,03,996.80	
E	15.00	16.50	25	110.00	375	41,250	
F	9.00	27.00	30	300.00	270	81.00	
				2605	3,70,246.8		

$$\text{Cost of living index} = \frac{\sum PW}{\sum W} = \frac{3,70,246.8}{2605} = 142.123$$

The small difference observed between the index by the Aggregative Method (142.13) and the index by the Family Budget Method (142.123) is due to the approximation in the value of price relatives ($= 108.33$) in commodity D.

13.14 PROBLEMS OF INDEX NUMBER CONSTRUCTION

Following four factors require careful consideration when planning the construction of indexes.

- Choice of base period
- Choice of weights
- Selection of the items to include
- purpose (scope) of the index

Choice of Base Period The base period associated with an index number is a period (or year) that is used as a basis for comparing percentage changes in prices or quantities in a given period. No matter what period of time is used as base period, the value of the index number for this period is 100. The period chosen should be one that corresponds to a time from which the relative changes are to be measured.

The following two considerations may be kept in view while deciding the base period:

- (i) The period chosen as a base should be fairly recent so that comparisons are not excessively affected by changing technology, product quality, purchasing habits etc.
- (ii) The base period should be a normal period, that is, it should not reflect a period in which fluctuations in price or quantity figures are either too low or too high. If the base period selected falls in an excessive (severely depressed) economic activities, then all indexes will appear to indicate poor (good) performance relative to this period.

Choice of Weights The weights used while computing an index reflect the relative importance of the individual items or commodities. The problem faced in index number construction is to decide *typical quantities* (consumption levels) and *prices* to compute *value* which measures the relative importance of items. The weights must be periodically revised in order to reflect the current behaviour of fluctuation in price and quantity.

Selection of Items The items or commodities to be included in the construction of an index should be carefully selected. Only those commodities should be included which would make the index most representative. When a large number of commodities are included in the construction of an index or all the necessary data are not available, then a *judgement sampling* tends to be used in preference to *simple random sampling to ensure representativeness for the purpose of the index*.

While selecting items to include in the construction of index it is important to decide which items *best relate to the purpose of index* and which set of products best represent a given item (e.g. cereals are included in a consumer price index, but which brand of cereal and what quantity?)

Purpose (Scope) of the Index Since different types of indexes available serve different purposes when applied to the same data, therefore the choice of an appropriate index depends on

- (i) the amount of accuracy required
- (ii) the frequency of measurement
- (iii) the choice of base period and
- (iv) the choice of the item.

If the mix of items is not readily measurable in the time period or if it is not possible to achieve accuracy from the data available, then it is advised to rethink of the purpose of constructing an index.

Since quality, accuracy, reliability, and adequacy of the data play an important role in the construction of an index, therefore data should be collected from a reliable source. The type of data required largely depends on the purpose of an index number. The purpose, once defined, helps to determine the data source availability.

While constructing an index, a choice has to be made between arithmetic mean and geometric mean. The use of A.M. gives more weights to commodities with high prices (though less important) and less weight to low-priced commodities. The G.M. is considered as the best average in the construction of an index due to the following reasons:

- (i) The G.M. gives equal weights to equal ratio or relative changes
- (ii) Major variations in the values of individual commodities do not influence the G.M.
- (iii) Index numbers calculated using G.M. are reversible and therefore base shifting is possible

Conceptual Questions 13B

11. (a) Discuss the various problems faced in the construction of index numbers.
 (b) Explain the problem faced in the construction of cost of living index.
 12. Discuss the importance and use of weights in the construction of general price index numbers.
 13. What is Fisher's Ideal index? Why is it called ideal? Show that it satisfies both the time reversal test as well as the factor reversal test. [Sukhadia Univ., MBA, 1998]
 14. Laspeyre's price index generally shows an upward trend in the price changes while Paasche's method shows a downward trend on them. Elucidate the statement. [Delhi Univ., MBA, 1997]
 15. Explain the Time Reversal Test and Factor Reversal Test with the help of suitable examples. [Osmania Univ., MBA, 1998]
 16. Distinguish between deflating and splicing of index numbers. [CA, Nov., 1999]
 17. What is the cost of living index number? Is it the same as the consumer price index number?
 18. What is the chain base method of construction of index numbers and how does it differ from the fixed base method?
 19. It is said that index numbers are a specialized type of averages. How far do you agree with this statement?
- Explain briefly the Time Reversal and Factor Reversal Tests. [Osmania Univ., MBA, 1995]
20. What are the Factor Reversal and Circular tests of consistency in the selection of an appropriate index formula? Verify whether Fisher's Ideal Index satisfies such tests. [CA May, 1996]
 21. What is the major difference between a weighted aggregate index and a weighted average of relatives index?
 22. What are the tests to be satisfied by a good index number? Examine how far they are met by Fisher's Ideal index number. [CA, May, 1996]
 23. Define Laspeyre's and Paasche's index numbers. It is said that Laspeyre's price Index tends to overestimate price changes while Paasche's price index tends to underestimate them. Put forward a possible explanation to substantiate this statement.
 24. What weights are used in Laspeyre's, Paasche's, and Marshall-Edgeworth's price index numbers? Prove that the Marshall-Edgeworth's price index number lies between the Paasche's and Laspeyre's price indexes.
 25. Theoretically geometric mean is the best average in the construction of index numbers but in practice mostly arithmetic mean is used. Discuss.
 26. What are the tests prescribed for a good index number? Describe the index number which satisfies these tests.

Formulae Used

$$1. \text{ Price relatives in period } n, P_{0n} = \frac{p_n}{p_0} \times 100$$

$$\text{Quantity relative in period } n, Q_{0n} = \frac{q_n}{q_0} \times 100$$

$$\text{Value relative in period } n, V_{0n} = \frac{\sum p_n q_n}{\sum p_0 q_0} \times 100$$

$$2. \text{ Unweighted aggregate price index in period } n$$

$$P_{0n} = \frac{\sum p_n}{\sum p_0} \times 100$$

Simple average of price relative

$$P_{0n} = \frac{1}{2} \sum \left(\frac{p_n}{p_0} \right) \times 100$$

Simple G.M. of price relative

$$P_{0n} = \text{antilog} \left[\frac{1}{n} \sum \left(\frac{p_n}{p_0} \right) \times 100 \right]$$

Simple aggregate quantity index

$$Q_{0n} = \frac{\sum q_n}{\sum q_0} \times 100$$

$$3. \text{ Weighted aggregate price indexes}$$

(a) Weighted aggregate method in period n

$$P_{0n} = \frac{\sum p_n q}{\sum p_0 q} \times 100$$

$$\text{Laspeyre's index, } I_p(L) = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100$$

$$\text{Paasche's index, } I_p(P) = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100$$

Marshall-Edgeworth's index

$$I_p(M-E) = \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)} \times 100$$

Dorbish and Bowley's index

$$I_p(D-B) = \frac{1}{2}(L + P) \times 100$$

$$\text{Fisher's ideal index, } = \sqrt{L \times P} \times 100$$

(b) Weighted average of price relatives in period n

$$P_{0n} = \frac{\sum \left(\frac{p_n}{p_0} \times 100 \right) W}{\sum W}$$

Weighted average of price relatives

$$P_{0n} = \frac{\sum \left(\frac{p_1}{p_0} \times 100 \right) (p_0 q_0)}{\sum p_0 q_0}$$

(base year value as weights)

Weighted average of price relatives

$$P_{0n} = \frac{\sum \left(\frac{p_1}{p_0} \times 100 \right) (p_1 q_1)}{\sum p_1 q_1}$$

(current year value as weights)

4. Quantity indexes

(a) Unweighted quantity index in period n

$$Q_{0n} = \frac{\sum q_n}{\sum q_0} \times 100$$

Simple average of quantity relative

$$Q_{0n} = \frac{1}{n} \sum \left(\frac{q_n}{q_0} \times 100 \right)$$

(b) Weighted quantity index in period n

$$Q_{0n} = \frac{\sum q_n W}{\sum q_0 W} \times 100$$

5. Tests for adequacy or consistency

Time reversal test: $P_{0n} \times P_{n0} = 1$

Factor reversal test: $P_{0n} \times Q_{0n} = \frac{\sum p_n q_n}{\sum p_0 q_0}$

Circular test: $P_{01} \times P_{12} \times P_{23} \times \dots \times P_{(n-1)n} \times P_{n0} = 1$

6. Link relative = $\frac{\text{Current period price}}{\text{Price of the preceding period}} \times 100$

Chain index = $\frac{\text{Current period's link relative} \times \text{Preceding period's chain index}}{100}$

Review Self-Practice Problems

- 13.29** Construct an index number for each year from the following average annual price of cotton with 1989 as the base year.

Year	Price (Rs)	Year	Price (Rs)
1989	75	1994	70
1990	50	1995	69
1991	65	1996	75
1992	60	1997	84
1993	72	1998	80

- 13.30** Compute a price index for the following by (a) simple aggregative method and (b) average of price relatives method by using both arithmetic and geometric mean.

Commodity	:	A	B	C	D	E	F
Price (Rs/unit) in 1991 :		20	30	10	25	40	50
Price (Rs/unit) in 2001 :		25	30	15	35	45	55

- 13.31** From the following average of three groups of commodities, find out index number, using arithmetic and geometric mean.

Group	1971	1992	1993	1994
A	8	12	16	20
B	32	40	48	60
C	16	20	32	40

- 13.32** The price quotations of four different commodities for 1996 and 1997 are given below. Calculate the index number for 1997 with 1996 as base by using (i) the simple average of price relatives and (ii) the weighted average of price relatives.

Comm	Unit	Weight (Rs 1000)	Price (in Rs per unit)	
			1996	1997
A	Kg	5	2.00	4.50
B	Quintal	7	2.50	3.20
C	Dozen	6	3.00	4.50
D	Kg	2	1.00	1.80

- 13.33** From the chain base index numbers given below, find the fixed base index numbers.

Year	:	1996	1997	1998	1999	2000
Chain base index :		80	110	120	90	140

- 13.34** The following are the group index numbers and the group weights of an average working class family's budget. Construct the cost of living number.

Group	Index Number	Weight
Food	330	50
Clothing	208	10
Fuel and lighting	200	12
House rent	162	12
Miscellaneous	180	16

- 13.35** In 1988, for working class people, wheat was selling at an average price of Rs 120 per 20 kg. Cloth Rs 20 per metre, house rent Rs 300 per house and other items Rs 100 per unit. By 1998 cost of wheat rose by Rs 160 per 20 kg, rent by Rs 450 house and other items doubled in price. The working class cost of living index for the year 1998 with 1988 as base was 160. By how much did the price of cloth rise during the period?

- 13.36** Calculate the cost of living index from the following data:

Items	Quantity Consumed per Year in the Given Year	Price (in Rs per Unit) in the Base Year	Given Year
Rice (qt)	2.50×12	12	25
Pulses (kg)	3×12	4	0.6
Oil (litre)	2×12	1.5	2.2
Clothing (metres)	6×12	0.75	1.0
Housing (per month)	—	20	30
Miscellaneous (per month)	—	10	15

- 13.37** Compute the Consumer Price Index number from the following:

Group	Base Year Price (Rs)	Current Year Price (Rs)	Weight (Per cent)
Food	400	550	35
Rent	250	300	25
Clothing	500	600	15
Fuel	200	350	20
Entertainment	150	225	5

[Mangalore Univ., BCom, 1997]

- 13.38** In calculating a certain cost of living index number, the following weights were used: Food 15, Clothing 3, Rent 4, Fuel and Light 2, Miscellaneous 1. Calculate the index for the period when the average percentage increases in prices of items in the various groups over the base

period were 32, 54, 47, 78, and 58 respectively.

Suppose a business executive was earning Rs 2050 in the base period, what should be his salary in the current period if his standard of living is to remain the same?

[Bangalore Univ., BCom, 1999]

- 13.39** Construct the cost of living index number from the following data:

Group	Weights	Group Index
Food	47	247
Fuel and Lighting	7	293
Clothing	8	289
House rent	13	100
Miscellaneous	14	236

[Vikram Univ., MBA, 1996]

- 13.40** During a certain period the cost of living index goes up from 110 to 200 and the salary of a worker is also raised from Rs 3250 to Rs 5000. Does the worker really gain, and if so, by how much in real terms?

- 13.41** An enquiry into the budgets of middle class families in a certain city gave the following information.

Expenses	Food 35%	Fuel 10%	Clothing 20%	Rent 15%	Miscellaneous 20%
Prices (Rs) 1990 : 150	25	75	30	40	
Prices (Rs) 1991 : 145	23	65	30	45	

What is the Cost of Living Index number of 1991 as compared with that of 1990?

Hints and Answers

13.29	Year	Price	Index Number (Base 1989)
	1989	75	100
	1990	50	$\frac{50}{75} \times 100 = 66.67$
	1991	65	$\frac{65}{75} \times 100 = 86.57$
	1992	60	$\frac{60}{75} \times 100 = 80.00$
	1993	72	$\frac{72}{75} \times 100 = 96.00$
	1994	70	$\frac{70}{75} \times 100 = 93.33$
	1995	69	$\frac{69}{75} \times 100 = 92.00$
	1996	75	$\frac{75}{75} \times 100 = 100.00$
	1997	84	$\frac{84}{75} \times 100 = 112.00$
	1998	80	$\frac{80}{75} \times 100 = 106.67$

- 13.30**

Commodity	p_0	p_1	$P = \frac{p_1}{p_0} \times 100$	Log P
A	20	25	125	2.0969
B	30	30	100	2.0000
C	10	15	150	2.1761
D	25	35	140	2.1461
E	40	45	112.5	2.0511
F	50	55	110	2.0414
	175	205	737.5	12.5116

$$\text{Simple aggregative index} = \frac{\sum p_1}{\sum p_0} \times 100 = \frac{205}{175} \times 100 = 117.143$$

$$\text{Arithmetic mean of price relatives} = \frac{\sum P}{n} = \frac{737.5}{6} = 122.92$$

$$\text{Geometric mean of price relatives} = \text{antilog} \left[\frac{\sum \log P}{n} \right] = \text{antilog} \left(\frac{12.5116}{6} \right) = 121.7$$

13.31

Group	1991		1992		1993		1994	
	Price	Relative	Price	Relative	Price	Relative	Price	Relative
A	8	100	12	150	16	200	20	250
B	32	100	40	125	48	150	60	187.5
C	16	100	20	125	32	200	40	250
Total		300		400		550		687.5
Mean of price relatives		100		133.3		183.3		229.2
G.M. of price relatives		100		132.83		181.66		227.15

13.32

Comm	Unit	Weight	Price	$P = \frac{p_1}{p_0} \times 100$		PW	
				(Rs, 1000)			
				1996	1997		
				W	p_0	p_1	
A	Kg	5	2.00	4.50	225	1,125	
B	Quint	7	2.50	3.20	128	896	
C	Dozen	6	3.00	4.50	150	900	
D	Kg	2	1.00	1.80	180	360	
		20			283	3,281	

(i) Simple average of price relatives

$$P_{01} = \frac{1}{n} \sum \left(\frac{p_1}{p_0} \times 100 \right) = \frac{683}{4} = 170.75$$

(ii) Weighted average of price relatives method:

$$P_{01} = \frac{\Sigma PW}{\Sigma W} = \frac{3281}{20} = 164.05$$

13.33 The formula for converting a Chain Base Index (CBI) number to a Fixed Base Index (FBI) Number is
Current years FBI

$$= \frac{\text{Current years CBI} \times \text{Previous years FBI}}{100}$$

Conversion of CBI to FBI

Year	Chain Base Index	Fixed Base Index
1996	80	80
1997	110	$\frac{110 \times 80}{100} = 88$
1998	120	$\frac{120 \times 88}{100} = 105.60$
1999	90	$\frac{90 \times 105.6}{100} = 95.04$
2000	140	$\frac{140 \times 95.04}{100} = 133.06$

13.34

Groups	Index No. P	Weights W	PW
Food	330	50	16,500
Clothing	208	10	2080
Fuel and Lighting	200	12	2400
House Rent	162	12	1944
Miscellaneous	180	16	2880
		100	25,804

$$\text{Cost of living index} = \frac{\Sigma PW}{\Sigma W} = \frac{25,804}{100} = 258.04$$

13.35

Commodity	Price	Index No. 1998	Price	Index No. 1998
Wheat	120	100	160	$\frac{160}{120} \times 100 = 150$
Cloth	20	100	x	$\frac{x}{20} \times 100 = 5x$
House rent	300	100	450	$\frac{450}{300} \times 100 = 150$
Miscell	100	100	200	$\frac{200}{100} \times 100 = 200$
				500 + 5x

The index for 1998 is 160. Thus the sum of the index numbers of the four commodities would be $160 \times 4 = 640$. Hence, $500 + 5x = 640$ or $x = 28$. The rise in the price of cloth was Rs 8 per metre.

13.36

Items	Quantity				
	Consumed q ₁	p_0	p_1	$p_1 q_1$	$p_0 q_1$
Rice (qtl)	2.50×12	12.00	25.0	750.0	360.0
Pulses (kg)	3×12	0.40	0.6	21.6	14.0
Oil (litres)	2×12	1.50	2.2	52.8	36.0
Clothing (mt)	6×12	0.75	1.0	72.0	54.0
Housing (per month)	—	20	30	360.0	240.0
Miscellaneous (per month)	—	10	15	$\frac{180.0}{120.0}$	$\frac{120.0}{824.4}$

$$\text{Cost of living index} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{1436.4}{824.4} \times 100 = 174.24$$

13.37

Group	p_0	q_1	$P = \frac{p_1}{q_0} \times 100$	Weight	PW
<i>W</i>					
Food	400	550	137.5	35	4812.5
Rent	250	300	120.0	25	3000.0
Clothing	500	600	120.0	15	1800.0
Fuel	200	350	175.0	20	3500.0
Entertainment	150	225	150.0	5	750.0
Total				100	13,862.5

$$\text{Consumer price index} = \frac{\Sigma PW}{\Sigma W} = \frac{13,862.5}{100} = 138.63$$

13.38

Group	Average Per cent	Group Index Increase in P	Price Weight	PW
			<i>W</i>	
Food	32	132	15	1980
Clothing	54	154	3	462
Rent	47	147	4	588
Fuel and light	78	178	2	356
Miscellaneous	58	158	1	158
Total			25	2544

$$\text{Cost of living index} = \frac{\Sigma PW}{\Sigma W} = \frac{3544}{25} = 141.76$$

For maintaining the same standard, the business executive should get $\frac{2050 \times 141.76}{100} = \text{Rs } 2906.08$.

13.39

Group	Weights (W)	Group Index (P)	PW
Food	47	247	11609
Fuel and lighting	7	293	2051
Clothing	8	289	2312
House rent	13	100	1300
Miscellaneous	14	236	3304
Total	89		20,576

$$\text{Cost of living} = \frac{\Sigma PW}{\Sigma W} = \frac{20,576}{89} = 231.19$$

$$\begin{aligned}\text{13.40 Real wage of Rs } 3250 &= \frac{\text{Actual wage}}{\text{Cost of living index}} \times 100 \\ &= \frac{3250}{110} \times 100 = \text{Rs } 2954.54\end{aligned}$$

Real wage of Rs 5000 = $\frac{5000}{200} \times 100 = \text{Rs } 2500$
which is less than Rs 2954.54

Since the real wage of Rs 5000 is less than that of Rs 3250, the worker does not really gain, real wage decrease by Rs $(2954.54 - 2500) = \text{Rs } 45.45$.

13.41

Expenses on	p_0	p_1	$P = \frac{p_1}{q_0} \times 100$	W	PW
Food	150	145	96.67	35	3383.45
Fuel	25	23	92.00	10	920.00
Clothing	75	65	89.67	20	1733.40
Rent	30	30	100.00	15	1500.00
Miscell	40	45	112.50	20	2250.00
Total				100	9786.85

$$\begin{aligned}\text{Cost of living index for 1991} &= \frac{\Sigma PW}{\Sigma W} = \frac{9786.85}{100} \\ &= 97.86.\end{aligned}$$

This page is intentionally left blank.

While an individual is an insolvable puzzle, in an aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to.

—Arthur Conan Doyle

Skewness, Moments, and Kurtosis

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- know the complementary relationship of skewness with measures of central tendency and dispersion in describing a set of data.
- understand 'moments' as a convenient and unifying method for summarizing several descriptive statistical measures.

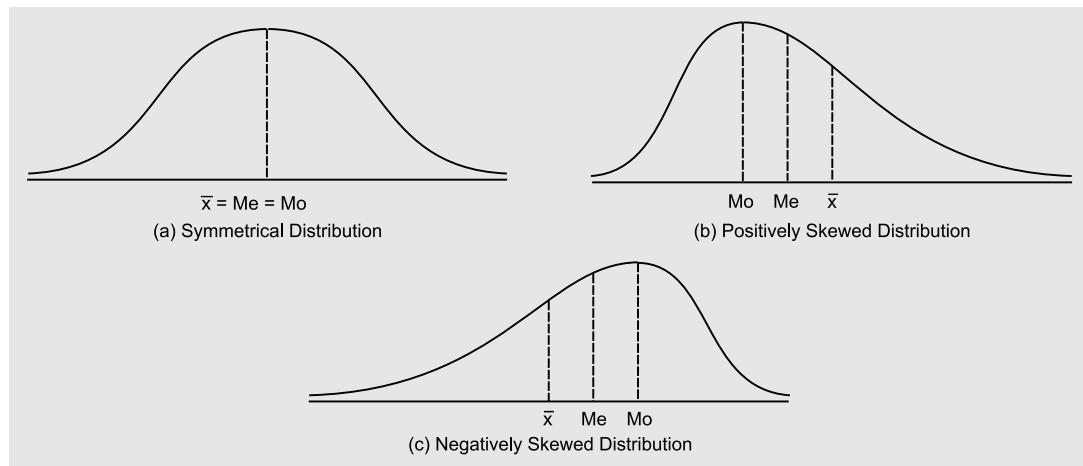
14.1 INTRODUCTION

In Chapter 4 we discussed measures of variation (or dispersion) to describe the spread of individual values in a data set around a central value. Such descriptive analysis of a frequency distribution remains incomplete until we measure the degree to which these individual values in the data set deviate from symmetry on both sides of the central value and the direction in which these are distributed. This analysis is important due to the fact that data sets may have the same mean and standard deviation but the frequency curves may differ in their shape. A frequency distribution of the set of values that is not 'symmetrical (normal)' is called *asymmetrical* or *skewed*. In a skewed distribution, extreme values in a data set move towards one side or tail of a distribution, thereby lengthening that tail. When extreme values move towards the upper or right tail, the distribution is positively skewed. When such values move towards the lower or left tail, the distribution is negatively skewed. As discussed, the mean, median, and mode are affected by the high-valued observations in any data set. Among these measures of central tendency, the mean value gets affected largely due to the presence of high-valued observations in one tail of a distribution. The mean value shifted substantially in the direction of high-values. The mode value is unaffected, while the median value, which is affected by the numbers but not the values of such observations, is also shifted in the direction of high-valued observations, but not as far as the mean. The median value changes about 2/3 as far as the mean value in the direction of high-valued observations (called extremes). Symmetrical and skewed distributions are shown in Fig. 14.1.

For a positively skewed distribution $A.M. > \text{Median} > \text{Mode}$, and for a negatively skewed distribution $A.M. < \text{Median} < \text{Mode}$. The relationship between these measures of central tendency is used to develop a **measure of skewness** called the *coefficient of skewness* to understand the degree to which these three measures differ.

Measure of skewness: The statistical technique to indicate the direction and extent of skewness in the distribution of numerical values in the data set.

Figure 14.1
Comparison of Three Data Sets Differing in Shape



From the above discussion, two points of difference emerge between variation and skewness:

- (i) Variation indicates the amount of spread or dispersion of individual values in a data set around a central value, while skewness indicates the direction of dispersion, that is, away from symmetry.
- (ii) Variation is helpful in finding out the extent of variation among individual values in a data set, while skewness gives an understanding about the concentration of higher or lower values around the mean value.

14.2 MEASURES OF SKEWNESS

The degree of skewness in a distribution can be measured both in the *absolute* and *relative* sense. For an asymmetrical distribution, the distance between mean and mode may be used to measure the degree of skewness because the mean is equal to mode in a symmetrical distribution. Thus,

$$\begin{aligned} \text{Absolute } S_k &= \text{Mean} - \text{Mode} \\ &= Q_3 + Q_1 - 2 \text{ Median} \text{ (if measured in terms of quartiles).} \end{aligned}$$

For a positively skewed distribution, Mean > Mode and therefore S_k is a positive value, otherwise it is a negative value. This difference is taken to measure the degree of skewness because in an asymmetrical distribution, mean moves away from the mode. Larger the difference between mean and mode, whether positive or negative, more is the asymmetrical distribution or skewness. This difference, however, may not be desirable for the following reasons:

- (i) The difference between mean and mode is expressed in the same units as the distribution and therefore cannot be used for comparing skewness of two or more distributions having different units of measurement.
- (ii) The difference between mean and mode may be large in one distribution and small in another, although the shape of their frequency curves is the same.

In order to overcome these two shortcomings and to make valid comparisons between skewness of two or more distributions, the absolute difference has to be expressed in relation to the standard deviation—a measure of dispersion. Since we want to express any measure of skewness as a pure (relative) number, therefore this distance is expressed in terms of the unit of measurement in units of the standard deviation.

14.2.1 Relative Measures of Skewness

The following are three important relative measures of skewness.

Karl Pearson's coefficient of skewness

The measure suggested by Karl Pearson for measuring coefficient of skewness is given by:

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} = \frac{\bar{x} - Mo}{\sigma} \quad (14-1)$$

where Sk_p = Karl Pearson's coefficient of skewness.

Since a mode does not always exist uniquely in a distribution, therefore it is convenient to define this measure using median. For a moderately skewed distribution the following relationship holds:

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

When this value of mode is substituted in the Eqn. (14-1) we get

$$Sk_p = \frac{3(\bar{x} - \text{Med})}{\sigma} \quad (14-2)$$

Theoretically, the value of Sk_p varies between ± 3 . But for a moderately skewed distribution, value of $Sk_p = \pm 1$. Karl Pearson's method of determining coefficient of skewness is particularly useful in open-end distributions.

Bowley's Coefficients of Skewness

The method suggested by Prof. Bowley is based on the relative positions of the median and the quartiles in a distribution. If a distribution is symmetrical, then Q_1 and Q_3 would be at equal distances from the value of the median, that is,

$$\text{Median} - Q_1 = Q_3 - \text{Median}$$

$$\text{or} \quad Q_3 + Q_1 - 2 \text{ Median} = 0 \quad \text{or} \quad \text{Median} = \frac{Q_3 + Q_1}{2}$$

This shows that the value of median is the mean value of Q_1 and Q_3 . Obviously in such a case, the absolute value of the coefficient of skewness will be zero.

When a distribution is asymmetrical, quartiles are not at equal distance from the median. The distribution is positively skewed, if $Q_1 - Me > Q_3 - Me$, otherwise negatively skewed.

The absolute measure of skewness is converted into a relative measure for comparing distributions expressed in different units of measurement. For this, absolute measure is divided by the inter-quartile range. That is,

$$\text{Relative } Sk_b = \frac{Q_3 + Q_1 - 2 \text{ Med}}{Q_3 - Q_1} = \frac{(Q_3 - \text{Med}) - (\text{Med} - Q_1)}{(Q_3 - \text{Med}) + (\text{Med} - Q_1)} \quad (14-3)$$

In a distribution, if $Med = Q_1$, then $Sk_b = \pm 1$, but if $Med = Q_3$ then $Sk_b = -1$. This shows that the value of Sk_b varies between ± 1 for moderately skewed distribution. This method of measuring skewness is quite useful in those cases where (i) mode is ill-defined and extreme observations are present in the data, (ii) the distribution has open-ended classes. These two advantages of Bowley's coefficient of skewness indicate that it is not affected by extreme observations in the data set.

Remark: The values of Sk_b obtained by Karl Pearson's and Bowley's methods cannot be compared. On certain occasions it is possible that one of them gives a positive value while the other gives a negative value.

Kelly's Coefficient of Skewness

The relative measure of skewness suggested by Prof. Kelly is based on percentiles and deciles:

$$Sk_k = \frac{P_{10} + P_{90} - 2P_{50}}{P_{90} - P_{10}} \quad \text{or} \quad \frac{D_1 + D_9 - 2D_5}{D_9 - D_1} \quad (14-4)$$

This method is an extension of Bowley's method in the sense that Bowley's method is based on the middle 50 per cent of the observations while this method is based on the observations between the 10th and 90th percentiles (or first and nineth deciles).

Example 14.1: Data of rejected items during a production process is as follows:

No of rejects : (per operator)	21–25	26–30	31–35	36–40	41–45	46–50	51–55
No. of operators :	5	15	28	42	15	12	3

Calculate the mean, standard deviation, and coefficient of skewness and comment on the results.

Solution: The calculations for mean, mode, and standard deviation are shown in Table 14.1

Table 14.1 Calculations for Mean, Mode and Standard Deviation

Class	Mid-value (m)	Frequency (f)	$d = \frac{m - A}{h} = \frac{m - 38}{5}$	fd	fd^2
21–25	23	5	-3	-15	45
26–30	28	15	-2	-30	60
31–35	33	28 $\leftarrow f_{m-1}$	-1	-28	28
36–40	38	42 $\leftarrow f_m$	0	0	0
41–45	43	15 $\leftarrow f_{m+1}$	1	15	15
46–50	48	12	2	24	48
51–55	53	3	3	9	27
		$N = 120$		-25	223

Let assumed mean, $A = 38$. Then

$$\bar{x} = A + \frac{\sum fd}{N} \times h = 38 - \frac{25}{120} \times 5 = 36.96 \text{ rejects per operator}$$

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times h \\ &= \sqrt{\frac{223}{120} - \left(\frac{-25}{120}\right)^2} \times 5 = 6.736 \text{ rejects per operator}\end{aligned}$$

By inspection, mode lies in the class 36–40. Thus

$$\begin{aligned}Mo &= l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h \\ &= 36 + \frac{42 - 28}{2 \times 42 - 28 - 15} = 36 + \frac{16}{41} \times 5 = 37.70\end{aligned}$$

$$\begin{aligned}Sk_p &= \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} = \frac{\bar{x} - Mo}{\sigma} \\ &= \frac{36.96 - 37.70}{6.73} = \frac{-0.74}{6.736} = -0.109\end{aligned}$$

Since the coefficient of skewness, $S_k = -0.109$, the distribution is skewed to the left (negatively skewed). Thus, the concentration of the rejects per operator is more on the lower values of the distribution to the extent of 10.9 per cent.

Example 14.2: The following is the information about the settlement of an industrial dispute in a factory. Comment on the gains and losses from the point of view of workers and that of management:

	Before	After
No. of Workers	3000	2900
Mean wages (Rs)	2200	2300
Median wages (Rs)	2500	2400
Standard deviation	300	260

Solution: The comments on gains and losses from both worker's and management's point of view are as follows:

Total Wages Bill

<i>Before</i>	<i>After</i>
$3000 \times 2200 = 66,00,000$	$2900 \times 2300 = 66,70,000$

The total wage bill has increased after the settlement of dispute, workers retained after the settlement are 50 workers less than the previous number.

After the settlement of dispute, the workers as a group are better off in terms of monetary gain. If the workers' efficiency remain same, then it is against the interest of management. But if the workers feel motivated, resulting in increased efficiency, then management can achieve higher productivity. This would be an indirect gain to management also.

Since workers retained after the settlement of dispute are less than the number employed before, it is against the interest of the workers.

Median Wages

The median wage after the settlement of dispute has come down from Rs 2500 to Rs 2400. This indicates that before the settlement, 50 per cent of the workers were getting wages above Rs 2500 but after the settlement, they will be getting only Rs 2400. It has certainly gone against the interest of the workers.

Uniformity in the Wage Structure

The extent of relative uniformity in the wage structure before and after the settlement can be determined by comparing the coefficient of variation as follows:

	<i>Before</i>	<i>After</i>
Coefficient of variation (CV)	$\frac{300}{2200} \times 100 = 13.63$	$\frac{260}{2300} \times 100 = 11.30$

Since CV has decreased after the settlement from 13.63 to 11.30, the distribution of wages is more uniform after the settlement, that is, there is now comparatively less disparity in the wages received by the workers. Such a position is good for both the workers and the management in maintaining a cordial work environment.

Pattern of the Wage Structure

The nature and pattern of the wage structure before and after the settlement can be determined by comparing the coefficients of skewness.

	<i>Before</i>	<i>After</i>
Coefficient of skewness, Sk_p	$\frac{3(2200 - 2500)}{300} = -3$	$\frac{3(2300 - 2400)}{260} = -1.15$

Since coefficient of skewness is negative and has increased after the settlement, therefore it suggests that number of workers getting low wages has increased and that of workers getting high wages has decreased after the settlement.

Example 14.3: From the following data on age of employees, calculate the coefficient of skewness and comment on the result

Age below (years) :	25	30	35	40	45	50	55
Number of employees :	8	20	40	65	80	92	100

[Delhi Univ., MBA, 1997]

Solution: The data are given in a cumulative frequency distribution form. So, to calculate the coefficient of skewness, convert this data into a simple frequency distribution as shown in Table 14.2.

Table 14.2 Calculations for Coefficient of Skewness

Age (years)	Mid-value (m)	Number of Employees (f)	$d = (m - A)/h$ $= (m - 37.5)$	fd	fd^2
20–25	22.5	8	-3	-24	72
25–30	27.5	12	-2	-24	48
30–35	32.5	20 $\leftarrow f_{m-1}$	-1	-20	20
35–40	37.5	25 $\leftarrow f_m$	0	0	0
40–45	42.5	1 $\leftarrow f_{m+1}$	1	15	15
45–50	47.5	12	2	24	48
50–55	52.5	8	3	24	72
$N = 100$				-5	275

$$\text{Mean, } \bar{x} = A + \frac{\sum fd}{N} \times h = 37.5 - \frac{5}{100} \times 5 = 37.25$$

Mode value lies in the class interval 35–40. Thus

$$\begin{aligned} Mo &= l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h \\ &= 35 + \frac{25 - 20}{2 \times 25 - 20 - 15} \times 5 = 35 + \frac{5}{15} \times 5 = 36.67 \end{aligned}$$

$$\begin{aligned} \text{Standard deviation, } \sigma &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times h \\ &= \sqrt{\frac{275}{100} - \left(\frac{-5}{100}\right)^2} \times 5 = \sqrt{2.75 - 0.0025} \times 5 = 8.29 \end{aligned}$$

Karl Pearson's coefficient of skewness:

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\sigma} = \frac{37.25 - 36.67}{8.29} = \frac{0.58}{8.29} = 0.07$$

The positive value of Sk_p indicates that the distribution is slightly positively skewed.

Example 14.4: (a) The sum of 50 observations is 500, its sum of squares is 6000 and median 12. Find the coefficient of variation and coefficient of skewness.

(b) For a moderately skewed distribution, the arithmetic mean is 100 and coefficient of variation 35, and Pearson's coefficient of skewness is 0.2. Find the mode and the median.

Solution: (a) Given that $N = 50$, $\Sigma x = 500$, $\Sigma x^2 = 6000$ and $Me = 12$.

$$\text{Mean, } \bar{x} = \frac{\Sigma x}{N} = \frac{500}{50} = 10$$

$$\text{Standard deviation, } \sigma = \sqrt{\frac{\Sigma x^2}{N} - (\bar{x})^2} = \sqrt{\frac{6000}{50} - (10)^2} = \sqrt{120 - 100} = 4.472$$

$$\text{Coefficient of variation, } CV = \frac{\sigma}{\bar{x}} \times 100 = \frac{4.472}{10} \times 100 = 44.7 \text{ per cent}$$

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean} = 3 \times 12 - 2 \times 10 = 16$$

$$\text{Coefficient of skewness, } Sk_p = \frac{\bar{x} - Mo}{\sigma} = \frac{10 - 16}{4.472} = -1.341$$

(b) Given that $\bar{x} = 100$, $CV = 35$, $Sk_p = 0.2$.

$$CV = \frac{\sigma}{\bar{x}} \times 100 \quad \text{or} \quad 35 = \frac{\sigma}{100} \times 100 \quad \text{or} \quad \sigma = 35$$

$$\text{Also } Sk_p = \frac{\bar{x} - Mo}{\sigma} \text{ or } 0.2 = \frac{100 - Mo}{35} \text{ or } Mo = 93$$

Mode = $3\text{Med} - 2\bar{x}$ or $93 = 3\text{Med} - 2 \times 100$ or $\text{Med} = 97.7$

Hence, Mode is 93 and median is 97.7.

Example 14.5: The data on the profits (in Rs lakh) earned by 60 companies is as follows:

Profits	:	Below 10	10–20	20–30	30–40	40–50	50 and above
No. of Companies	:	5	12	20	16	5	2

(a) Obtain the limits of profits of the central 50 per cent companies.

(b) Calculate Bowley's coefficient of skewness.

Solution: (a) Calculations for different quartiles are shown in Table 14.3.

Table 14.3 Computation of Quartiles

Profits (in Rs lakh)	Frequency (f)	Cumulative Frequency (c.f.)
Below 10	5	5
10–20	12	17 $\leftarrow Q_1$ Class
20–30	20	37
30–40	16	53 $\leftarrow Q_3$ Class
40–50	5	58
50 and above	2	60
$N = 60$		

Q_1 = size of $(N/4)$ th observation = $(60/4)$ th = 15th observation. Thus, Q_1 lies in the class 10–20, and

$$\begin{aligned} Q_1 &= l + \left\{ \frac{(N/4) - cf}{f} \right\} \times h \\ &= 10 + \left\{ \frac{15 - 5}{12} \right\} \times 10 = 10 + 8.33 = 18.33 \text{ lakh} \end{aligned}$$

Q_3 = size of $(3N/4)$ th observation = 45th observation. Thus, Q_3 lies in the class 30–40, and

$$\begin{aligned} Q_3 &= l + \left\{ \frac{(3N/4) - cf}{f} \right\} \times h \\ &= 30 + \left\{ \frac{45 - 37}{16} \right\} \times 10 = 30 + 5 = 35 \text{ lakh} \end{aligned}$$

Hence, the profit of central 50 per cent companies lies between Rs 35 lakhs and Rs 18.33 lakhs.

$$\text{Coefficient of quartile deviation, Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{35 - 18.33}{35 + 18.33} = 0.313$$

(b) Median = size of $(N/2)$ th observation = 30th observation. Thus, median lies in the class 20–30, and

$$Me = l + \left\{ \frac{(N/2) - cf}{f} \right\} \times h = 20 + \left\{ \frac{30 - 17}{20} \right\} \times 10 = 20 + 6.5 = 26.5 \text{ lakh}$$

$$\text{Coefficient of skewness, } Sk_b = \frac{Q_3 + Q_1 - 2\text{Med}}{Q_3 - Q_1} = \frac{35 + 18.33 - 2(26.5)}{35 - 18.33} = 0.02$$

The positive value of Sk_b indicates that the distribution is positively skewed and therefore, there is a concentration of larger values on the right side of the distribution.

Example 14.6: Apply an appropriate measure of skewness to describe the following frequency distribution.

Age (yrs)	Number of Employees	Age (yrs)	Number of Employees
Below 20	13	35–40	112
20–25	29	40–45	94
25–30	46	45–50	45
30–35	60	50 and above	21

[Bharthidasan Univ., MBA, 2001]

Solution: Since given frequency distribution is an open-ended distribution, Bowley's method of calculating skewness should be more appropriate. Calculations are shown in Table 14.4.

Table 14.4 Calculations for Bowley's Coefficient of Skewness

Age (yrs)	Number of Employees (f)	Cumulative Frequency (cf)
Below 20	13	13
20–25	29	42
25–30	46	88
30–35	60	148 ← Q_1 class
35–40	112	260
40–45	94	354 ← Q_3 class
45–50	45	399
50 and above	21	420
$N = 420$		

Q_1 = size of $(N/4)$ th observation = $(420/4) = 105$ th observation. Thus, Q_1 lies in the class 30–35, and

$$Q_1 = l + \frac{(N/4) - cf}{f} \times h = 30 + \frac{105 - 88}{60} \times 5 = 30 + 1.42 = 31.42 \text{ years}$$

Q_3 = size of $(3N/4)$ th observation = $(3 \times 420/4) = 315$ th observation. Thus, Q_3 lies in the class 40–45, and

$$Q_3 = l + \frac{(3N/4) - cf}{f} \times h = 40 + \frac{315 - 260}{94} \times 5 = 40 + 2.93 = 42.93 \text{ years}$$

Median = size of $(N/2)$ th = $(420/2) = 210$ th observation. Thus, median lies in the class 35–40, and

$$Me = l + \frac{(N/2) - cf}{f} \times h = 35 + \frac{210 - 148}{112} \times 5 = 35 + 2.77 = 37.77 \text{ years}$$

$$\begin{aligned} \text{Coefficient of skewness, } Sk_b &= \frac{Q_3 + Q_1 - 2\text{Med}}{Q_3 - Q_1} = \frac{42.93 + 31.42 - 2 \times 37.77}{42.93 - 31.42} \\ &= -\frac{1.19}{11.51} = -0.103 \end{aligned}$$

The negative value of Sk_b indicates that the distribution is negatively skewed.

Conceptual Questions 14A

- Explain the meaning of skewness using sketches of frequency curves. State the different measures of skewness that are commonly used. How does skewness differ from dispersion?
- Measures of central tendency, dispersion, and skewness are complementary to each other in describing a frequency distribution. Elucidate.

3. Distinguish between Karl Pearson's and Bowley's measure of skewness. Which one of these would you prefer and why? [Delhi Univ., MBA, 2000]
4. Define and discuss the 'quartiles' of a distribution. How are they used for measuring dispersion and skewness and point out the various methods of measuring skewness.
5. Explain briefly the different methods of measuring skewness. [Kumaon Univ., MBA, 2000]
6. Define and discuss the 'quartiles' of a distribution. How are they used for measuring variation and skewness.
7. Distinguish between variation and skewness and point out the various methods of measuring skewness.
8. Briefly mention the tests which can be applied to determine the presence of skewness.
9. Explain the term 'skewness'. What purpose does a measure of skewness serve? Comment on some of the well known measures of skewness.

Self-Practice Problems 14A

- 14.1** The following data relate to the profits (in Rs '000) of 1,000 companies:

Profits :	100–120	120–140	140–160	160–180
	180–200	200–220	220–240	
No. of companies :	17	53	199	194
	327	208	2	

Calculate the coefficient of skewness and comment on its value. [MD Univ., MBA, 2001]

- 14.2** A survey was conducted by a manufacturing company to find out the maximum price at which people would be willing to buy its product. The following table gives the stated price (in rupees) by 100 persons:

Price :	2.80–2.90	2.90–3.00	3.00–3.10
	3.10–3.20	3.20–3.30	
No. of persons:	11	29	18
	27	15	

Calculate the coefficient of skewness and interpret its value.

- 14.3** Calculate coefficient of variation and Karl Pearson's coefficient of skewness from the data given below:

Marks (less than) :	20	40	60	80	100
No. of students :	18	40	70	90	100

- 14.4** The following table gives the length of the life (in hours) of 400 TV picture tubes:

Length of Life (in hours)	No. of Picture Tubes	Length of Life (in hours)	No. of Picture Tubes
4000–4199	12	5000–5199	55
4200–4399	30	5200–5399	36
4400–4599	65	5400–5599	25
4600–4799	78	5600–5799	9
4800–4999	90		

Compute the mean, standard deviation, and coefficient of skewness.

- 14.5** Calculate Karl Pearson's coefficient of skewness from the following data:

Profit (in Rs lakh) :	Below 20	40	60	80	100
No. of companies :	8	20	50	64	70

- 14.6** From the following information, calculate Karl Pearson's coefficient of skewness.

Measure	Place A	Place B
Mean	256.5	240.8
Median	201.0	201.6
S.D.	215.0	181.0

- 14.7** From the following data, calculate Karl Pearson's coefficient of skewness:

Marks (more than) :	0	10	20	30	40	50	60	70	80
No. of students :	150	140	100	80	80	70	30	14	0

- 14.8** The following information was collected before and after an industrial dispute:

	Before	After
No. of workers employed	515	509
Mean wages (Rs)	4900	5200
Median wages (Rs)	5280	5000
Variance of wages (Rs)	121	144

Comment on the gains or losses from the point of view of workers and that of the management.

- 14.9** Calculate Bowley's coefficient of skewness from the following data

Sales (in Rs lakh) :	Below 50	60	70	80	90
No. of companies :	8	20	40	65	80

[Delhi Univ., MBA, 1998, 2003]

- 14.10** The following table gives the distribution of weekly wages of 500 workers in a factory:

Weekly Wages (Rs)	No. of Workers
Below 200	10
200–250	25
250–300	145
300–350	220
350–400	70
400 and above	30

- (a) Obtain the limits of income of the central 50 per cent of the observed workers.

- (b) Calculate Bowley's coefficient of skewness.

- 14.11** Find Bowley's coefficient of skewness for the following frequency distribution

No. of children per family :	0	1	2	3	4
No. of families :	7	10	16	25	18

- 14.12** In a frequency distribution, the coefficient of skewness based on quartiles is 0.6. If the sum of the upper and the lower quartiles is 100 and the median is 38, find the value of the upper quartile.

- 14.13** Calculate Bowley's measure of skewness from the following data:

Payment of Commission	No. of Salesmen	Payment of Commission	No. of Salesmen
100–120	4	200–220	80
120–140	10	220–240	32
140–160	16	240–260	23
160–180	29	260–280	17
180–200	52	280–300	7

- 14.14** Compute the quartiles, median, and Bowley's coefficient of skewness:

Income (in Rs)	No. of families
Below 200	25
200–400	40
400–600	80
600–800	75
800–1000	20
1000 and above	16

Hints and Answers

$$14.1 \quad \bar{x} = A + \frac{\sum fd}{N} \times h = 170 + \frac{393}{1000} \times 20 = 177.86$$

$$\text{Mo} = \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h = 180 + \frac{233}{252} \times 20 = 190.55$$

$$\sigma = h \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = 20 \sqrt{\frac{1741}{1000} - \left(\frac{393}{1000}\right)^2} = 25.2$$

$$\text{Sk}_p = \frac{\bar{x} - \text{Mo}}{\sigma} = \frac{177.86 - 190.55}{25.2} = -0.5035$$

$$14.2 \quad \bar{x} = A + \frac{\sum fd}{N} \times h = 3.05 + \frac{6}{100} \times 0.1 = 3.056$$

$$\text{Mo} = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h = 2.9 + \frac{29 - 11}{2 \times 29 - 11 - 18} \times 0.1 = 2.962$$

$$\sigma = h \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = 0.1 \sqrt{\frac{160}{100} - \left(\frac{6}{100}\right)^2} = 0.1264$$

$$\text{Sk}_p = \frac{\bar{x} - \text{Mo}}{\sigma} = \frac{3.056 - 2.962}{0.1264} = 0.744$$

$$14.3 \quad \bar{x} = A + \frac{\sum fd}{N} \times h = 50 - \frac{18}{100} \times 20 = 46.4$$

$$\sigma = h \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = 20 \sqrt{\frac{154}{100} - \left(\frac{-18}{100}\right)^2} = 24.56$$

$$\text{C.V.} = \frac{\sigma}{\bar{x}} \times 100 = \frac{24.56}{46.4} \times 100 = 52.93$$

$$\text{Mo} = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h = 39.52;$$

$$\text{Sk}_p = 0.280$$

$$14.4 \quad \bar{x} = A + \frac{\sum fd}{N} \times h = 4899.5 - \frac{108}{400} \times 200 = 4845.5$$

$$\sigma = h \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = 200 \sqrt{\frac{1368}{400} - \left(\frac{-108}{400}\right)^2} = 365.9$$

Mo = Mode lies in the class 4800–4999; but real class interval is 4799.5–4999.5

$$\begin{aligned} &= l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h \\ &= 4799.5 + \frac{12}{180 - 78 - 55} \times 200 = 4850.56; \end{aligned}$$

$$\text{Sk}_p = -0.014$$

$$14.5 \quad \bar{x} = A + \frac{\sum fd}{N} \times h = 50 - \frac{2}{70} \times 20 = 49.43$$

$$\text{Mo} = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h = 40 + \frac{14}{12 + 14} \times 20 = 50.76$$

$$\sigma = h \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = 20 \sqrt{\frac{80}{70} - \left(\frac{-2}{70}\right)^2} = 21.64; \quad \text{Sk}_p = -0.061$$

14.6 $a \equiv \bar{x}$ and $z \equiv \text{Mo}$;

Place A : Mode = 3Med – 2 \bar{x} = 90;

$$\text{Sk}_p = \frac{266.5 - 90}{215} = 0.823$$

Place B : Mode = 3Med – 2 \bar{x} = 123.2;

$$\text{Sk}_p = \frac{240.8 - 123.2}{181} = 0.649$$

Marks	No. of Students
0–10	10
10–20	40
20–30	20
30–40	0
40–50	10
50–60	40
60–70	16
70–80	14

$$\bar{x} = 39.27; \sigma = 22.81;$$

$$Sk_p = \frac{3(\bar{x} - \text{Med})}{\sigma} = \frac{3(39.27 - 45)}{22.81} = -0.754.$$

- 14.9** Q_1 lies in the class 50–60; $Q_1 = 60$; Q_3 lies in the class 70–80; $Q_3 = 78$

Median ($= Q_2$) lies in the class 60–70; Med = 70

$$\text{Bowley's coeff. of } Sk_b = \frac{Q_3 + Q_1 - 2\text{Med}}{Q_3 - Q_1} = -0.111$$

- 14.10** $Q_1 = (80/4) = 20$ th observation lies in the class 250–300; $Q_1 = 281.03$; $Q_3 = (3 \times 80)/4 = 60$ th observation lies in the class 300–350; $Q_3 = 344.32$

Median lies in the class 300–350, Me = 315.9

Bowley's coeff. of $Sk_b = -0.111$ (negatively skewed distribution)

- 14.11** $Q_1 = \text{size of } \left(\frac{n+1}{4}\right)\text{th observation} = 24$ th observation = 2

$Q_3 = \text{size of } \frac{3(n+1)}{4}\text{th observation} = 72$ th

observation = 4

$Me = \text{size of } \left(\frac{n+1}{2}\right)\text{th observation} = 48$ th

observation; $Sk_b = \frac{4+2-2(3)}{4-2} = 0$

- 14.12** Given $Sk_b = 0.6$; $Q_1 + Q_3 = 100$; $Me = 38$; $Q_3 = ?$

$$Sk_b = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1} \text{ or } 0.6 = \frac{100 - 2(38)}{Q_3 - (100 - Q_3)} \text{ or}$$

$$Q_3 = 70$$

- 14.13** $Q_1 = (n/4)\text{th observation} = 67.5$ th observation lies in class 180–200; $Q_1 = 183.26$

$Q_3 = \left(\frac{3n}{4}\right)\text{th observation} = 202.5$ th observation lies in class 220–240; $Q_3 = 227.187$

$Me = (n/2)\text{th observation} = 135$ th observation lies in class 200–220; $Me = 206$

$$Sk_b = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1} = -0.035$$

14.3 MOMENTS

According to R. A. Fisher, ‘A quantity of data which by its mere bulk may be incapable of entering the mind is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.’

Among these ‘relatively few quantities’ are those which are known as **moments**. Two of them, the *mean* and the *variance*, have already been discussed and two other higher moments are in common use. The higher moments are basically used to describe the characteristic of populations rather than samples.

The measures of central tendency, variability and skewness which have been discussed to describe a frequency distribution, may be classified into two groups:

- (i) Percentile system, and
- (ii) Moment system

The *percentile system* includes measures like median, quartiles, deciles, percentiles, and so on. The value of these measures represents a given proportion of frequency distribution.

The *moment system* includes measures like mean, average deviation, standard deviation, and so on. The value of these measures is obtained by taking the deviation of individual observations from a given origin. The term ‘moment’ is used in physics and refers to the measure of a force which may generate rotation. The possibility of generating such a force depends upon (i) the amount of force needed and (ii) the distance from the origin of the point at which the force is applied. The term moment used in statistics is analogous to the term used in physics, where (i) size of class intervals represents the ‘force’ and (ii) deviation of mid-value of each class from an observation represents the distance.

While calculating moments, if deviations are taken from the actual mean, then such moments are denoted by the Greek letter μ (mu). On the other hand, if deviations are taken from some assumed mean (or arbitrary value other than zero), then moments are denoted by Greek letter ν (nu) or μ' .

Moments: Represent a convenient and unifying method for summarizing certain descriptive statistical measures.

14.3.1 Moments about Mean

Let x_1, x_2, \dots, x_n be the n observations in a data set with mean \bar{x} . Then the r th moment about the actual mean of a variable both for ungrouped and grouped data is given by:

For ungrouped data: $\mu_r = \frac{1}{n} \sum (x - \bar{x})^r ; r = 1, 2, 3, 4.$

For grouped data: $\mu_r = \frac{1}{n} \sum f(x - \bar{x})^r ; r = 1, 2, 3, 4; N = \sum f_i$

For different values of $r = 1, 2, 3, 4$, different moments can be obtained as shown below:

Ungrouped data: $\mu_1 = \frac{1}{n} \sum (x - \bar{x}) = 0; \mu_2 = \frac{1}{n} \sum (x - \bar{x})^2 = \sigma^2$ (variance)

$$\mu_3 = \frac{1}{n} \sum (x - \bar{x})^3; \quad \mu_4 = \frac{1}{n} \sum (x - \bar{x})^4$$

For grouped data: $\mu_1 = \frac{1}{n} \sum f(x - \bar{x}); \quad \mu_2 = \frac{1}{n} \sum f(x - \bar{x})^2$

$$\mu_3 = \frac{1}{n} \sum f(x - \bar{x})^3; \quad \mu_4 = \frac{1}{n} \sum f(x - \bar{x})^4$$

The *first moment*, μ_1 about origin gives the *mean* and is a measure of central tendency.

$$\mu_1 = \frac{1}{n} \sum (x - 0) = \frac{1}{n} \sum x \leftarrow \text{A.M.}$$

The *second moment*, μ_2 about the mean is known as *variance* and is a measure of dispersion.

$$\mu_2 = \frac{1}{n} \sum (x - \bar{x})^2 \leftarrow \text{Variance}$$

The *third moment*, μ_3 about the mean indicates the symmetry or asymmetry of the distribution; its value is zero for symmetrical distribution.

$$\mu_3 = \frac{1}{n} \sum (x - \bar{x})^3$$

The *fourth moment*, μ_4 about the mean is a measure of Kurtosis (or flatness) of the frequency curve.

$$\mu_4 = \frac{1}{n} \sum (x - \bar{x})^4 \leftarrow \text{Kurtosis}$$

14.3.2 Moments about Arbitrary Point

When actual mean is in fractions, moments are first calculated about an assumed mean, say A, and then are converted about the actual mean, as shown below:

$$\begin{aligned} \text{For grouped data: } \mu'_r &= \frac{1}{n} \sum f(x - A)^r; \quad r = 1, 2, 3, 4 \\ &= \frac{1}{n} \sum f d^2 \times h^2, \text{ where } d = \frac{x - A}{h} \text{ or } dh = x - A \end{aligned}$$

$$\text{For ungrouped data: } \mu'_r = \frac{1}{n} \sum (x - A)^r; \quad r = 1, 2, 3, 4$$

$$\text{For } r = 1, \text{ we have } \mu'_1 = \frac{1}{n} \sum (x - A) = \frac{1}{n} \sum x - A = \bar{x} - A$$

14.3.3 Moments about Zero or Origin

The moments about zero or origin are obtained as follows:

$$v_r = \frac{1}{n} \sum f x^r; \quad r = 1, 2, 3, 4$$

The relationship among moments about zero and other moments is as follows:

$$\begin{aligned} v_1 &= A + \mu'_1, & v_2 &= \mu_2 + (v_1)^2 \\ v_2 &= \mu_3 + 3v_1 v_2 - 2v_1^3, & v_4 &= \mu_4 + 4v_1 v_3 - 6v_1^2 v_2 + 3v_1^4 \end{aligned}$$

14.3.4 Relationship Between Central Moments and Moments about any Arbitrary Point

$$\mu_r = \frac{1}{n} \sum (x - \bar{x})^r = \frac{1}{n} \sum \{x - A - (\bar{x} - A)\}^r = \frac{1}{n} \sum (x - A - \mu'_1)^r; \quad \mu_1 = \bar{x} - A$$

$$\begin{aligned}
&= \frac{1}{n} \left[\sum (x - A)^r - {}^r C_1 \mu'_1 \sum (x - A)^{r-1} + {}^r C_2 (\mu'_1)^2 \sum (x - A)^{r-2} + \dots \right. \\
&\quad \left. + (-1)^r (\mu'_1)^r \right] \\
&= \mu'_r - {}^r C_1 \mu'_1 \mu'_{r-1} + {}^r C_2 (\mu'_1)^2 \mu'_{r-2} + \dots + (-1)^r (\mu'_1)^r \tag{14-5}
\end{aligned}$$

From Eqn. (14-5) for various values of r , we have

$$\begin{aligned}
\mu_1 &= \mu'_1; \quad r = 1 & \mu_2 &= \mu'_2 - (\mu'_1)^2; \quad r = 2 \\
\mu_3 &= \mu'_3 - 3\mu'_1 \mu'_2 + 2(\mu'_1)^3; \quad r = 3 \\
\mu_4 &= \mu'_4 - 4\mu'_1 \mu'_3 + 6\mu'_2 (\mu'_1)^2 - 3(\mu'_1)^3; \quad r = 4
\end{aligned}$$

14.3.5 Moments in Standard Units

When expressed in standard units, moments about the mean of the population are usually denoted by the Greek letter α (alpha). Thus

$$\begin{aligned}
\alpha_r &= \frac{1}{n} \sum f z^r = \frac{1}{n} \sum f \left(\frac{x - \mu}{\sigma} \right)^r \text{ by definite of } z = \frac{x - \mu}{\sigma} \\
&= \frac{1}{\sigma^r} \frac{1}{n} \sum f (x - \mu)^r = \frac{\mu_r}{\sigma^r}
\end{aligned}$$

Hence, $\alpha_1 = 0$, $\alpha_2 = 1$, $\alpha_3 = \frac{\mu_3}{\sigma^3}$ and $\alpha_4 = \frac{\mu_4}{\sigma^4}$ for $r = 1, 2, 3$, and 4 respectively.

In the notations used by Karl Pearson, we have

$$\begin{aligned}
(i) \quad \beta_1 \text{ (Beta one)} &= \alpha_3^2 = \frac{\mu_3^2}{\mu_2^3} \\
(ii) \quad \beta_2 \text{ (Beta two)} &= \alpha_4 = \frac{\mu_4}{\mu_2^2}
\end{aligned}$$

In the notations used by R. A. Fisher, we have

$$\begin{aligned}
(i) \quad \gamma_1 \text{ (Gamma one)} &= \alpha_3 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3} \\
(ii) \quad \gamma_2 \text{ (Gamma two)} &= \alpha_4 - 3 \text{ or } \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{\mu^4 - 3\mu_2^2}{\mu_2^2}
\end{aligned}$$

All these coefficients are pure numbers and are independent of whatever unit the variable x may be expressed in. The values of α_3 and α_4 depend on the shape of the frequency curve and, therefore, can be used to distinguish between different shapes. Thus α_3 or β_1 is a measure of asymmetry about the mean or skewness and $\beta_1 = 0$ for a symmetrical distribution. A curve with $\alpha_3 > 0$ is said to have positive skewness and one with $\alpha_3 < 0$, negative skewness. For most distributions, α_3 lies between -3 and 3 .

The coefficient of skewness in terms of moments is given by

$$Sk = \frac{(\beta_2 + 3)\sqrt{\beta_1}}{2(5\beta_2 - 6\beta_1 - 9)}$$

If $\beta_1 = 0$ or $\beta_2 = -3$, then skewness is zero. But $\beta_2 = \frac{\mu_4}{\mu_2^2}$ can not be zero and hence the only condition for skewness to be zero is $\beta_1 = 0$ and the coefficient has no sign.

14.3.6 Sheppard's Corrections for Moments

While calculating higher moments of grouped frequency distributions, it is assumed that frequencies are concentrated at the mid-values of class-intervals. However, it causes certain errors while calculating moments. W. E. Sheppard proved that, if

- (i) the frequency curve of the distribution is continuous,
- (ii) the frequency tapers off to zero at both ends, and
- (iii) the member of classes are not too large,

then the above assumption that frequencies are concentrated at the mid-value of the class intervals is corrected using Shppard's corrections as follows:

$$\mu_2 \text{ (corrected)} = \mu_2 - \frac{h^2}{12}$$

$$\mu_4 \text{ (corrected)} = \mu_4 - \frac{1}{2} h^2 \mu_2 + \frac{7}{240} h^4$$

where h is the width of the class interval, μ_1 and μ_3 need no correction.

Example 14.7: The first four moments of a distribution about the origin are 1, 4, 10, and 46 respectively. Obtain the various characteristics of the distribution on the basis of the information given. Comment upon the nature of the distribution.

Solution: In the usual notations, we have

$$A = 0, \mu'_1 = 1, \mu'_2 = 4, \mu'_3 = 10 \text{ and } \mu'_4 = 46$$

$$\bar{x} = \text{first moment about origin} = \mu'_1 = 1$$

$$\text{Variance} (\sigma^2) = \mu_2 = \mu'_2 - (\mu'_1)^2 = 4 - 1 = 3$$

$$\text{S.D.} (\sigma) = \sqrt{\sigma^2} = \sqrt{3} = 1.732$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3 = 10 - 3(4)(1) + 2(1)^3 = 0.$$

Karl Pearson's coefficient of skewness

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \quad \text{or} \quad \gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3}$$

Substituting values in the above formula, we get $\gamma_1 = 0$ ($\beta_1 = 0$). This shows that the given distribution is symmetrical, and hence Mean = Median = Mode for the given distribution.

Example 14.8: The first three moments of a distribution about the value 1 of the variable are 2, 25 and 80. Find the mean, standard deviation and the moment-measure of skewness.

Solution: From the data of the problem, we have

$$\mu'_1 = 2, \mu'_2 = 25, \mu'_3 = 80 \text{ and } A = 1.$$

The moments about the arbitrary point $A = 1$ are calculated as follows:

$$\text{Mean, } \bar{x} = \mu'_1 + A = 2 + 1 = 3$$

$$\text{Variance, } \mu_2 = \mu'_2 - (\mu'_1)^2 = 25 - (2)^2 = 21$$

$$\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2(\mu'_1)^3 = 80 - 3(2)(25) + 2(2)^2 = -54$$

$$\text{Standard deviation, } \sigma = \sqrt{\mu_2} = \sqrt{21} = 4.58$$

$$\text{Moment-measure of skewness, } \gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{-54}{(21)^{3/2}} = \frac{-54}{96.234} = -0.561$$

Example 14.9: Following is the data on daily earnings (in Rs) of employees in a company:

Earnings	:	50–70	70–90	90–110	110–130	130–150	150–170	170–190
No. of workers	:	4	8	12	20	6	7	3

Calculate the first four moments about the point 120. Convert the results into moments about the mean. Compute the value of γ_1 and γ_2 and comment on the result.

[Delhi Univ., MBA, 1990, 2002]

Solution: Calculations for first four moments are shown in Table 14.5:

Table 14.5 Computation of First Four Moments

Class	Mid-value	Frequency	$d = \frac{m - 120}{20}$	fd	fd^2	fd^3	fd^4
50–70	60	4	-3	-12	36	-108	324
70–90	80	8	-2	-16	32	-64	128
90–110	100	12	-1	-12	12	-12	12
110–130	120	20	0	0	0	0	0
130–150	140	6	1	6	6	6	6
150–170	160	7	2	14	28	56	112
170–190	180	3	3	9	27	81	243
		60		-11	141	-41	825

The moments about some arbitrary origin or point ($A = 120$) is given by

$$\begin{aligned}\mu'_r &= \left(\frac{1}{n}\right) \sum f(x - A)^r \quad (\text{for grouped data}) \\ &= \frac{1}{n} (\sum f d^r) h^r; \quad d = \frac{m - A}{h} \quad \text{or } m - A = hd\end{aligned}$$

For $A = 120$ and $x = m$, we get

$$\begin{aligned}\mu'_1 &= \frac{1}{n} \sum f d \times h = \frac{1}{60} (-11) \times 20 = -3.66 \\ \mu'_2 &= \frac{1}{n} \sum f d^2 \times h^2 = \frac{1}{60} (141) \times (20)^2 = 940 \\ \mu'_3 &= \frac{1}{n} \sum f d^3 \times h^3 = \frac{1}{60} (-41) (20)^3 = -5,466.66 \\ \mu'_4 &= \frac{1}{n} \sum f d^4 \times h^4 = \frac{1}{60} (825) (20)^4 = 22,00,000\end{aligned}$$

The moments about actual mean ($\mu'_2 = 940$) is given by

$$\begin{aligned}\mu_1 &= 0 \\ \mu_2 &= \mu'_2 - (\mu'_1)^2 = 940 - (-3.66)^2 = 926.55 \\ \mu_3 &= \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3 = -5,466.66 - 3(940)(-3.66) + 2(-3.66)^3 \\ &= -5,466.66 + 10,340.094 - 98.59 = 4,774.83 \\ \mu_4 &= \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2(\mu'_1)^2 - 3(\mu'_1)^4 \\ &= 22,00,000 - 4(-5,466.66)(-3.66) + 6(940)(-3.66)^2 - 3(-3.66)^4 \\ &= 22,00,000 - 80,178.50 + 7,582.03 - 542.27 = 21,95,107.20\end{aligned}$$

Since μ_3 is positive, therefore the given distribution is positively skewed. The relative measure of skewness is given by

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\mu_2 \sqrt{\mu_2}} = \frac{4774.83}{926.55 \sqrt{926.55}} = 0.169$$

Thus, $\beta_1 = \gamma_1^2 = 0.0285$. This implies that distribution is positively skewed to the right.

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{21,95,107.20}{(926.55)^2} = 2.56$$

$$\gamma_2 = \beta_2 - 3 = 2.56 - 3 = -0.44$$

Since γ_2 is negative, the distribution is platykurtic.

14.4 KURTOSIS

The measure of kurtosis, describes the degree of concentration of frequencies (observations) in a given distribution. That is, whether the observed values are concentrated more around the mode (a peaked curve) or away from the mode towards both tails of the frequency curve.

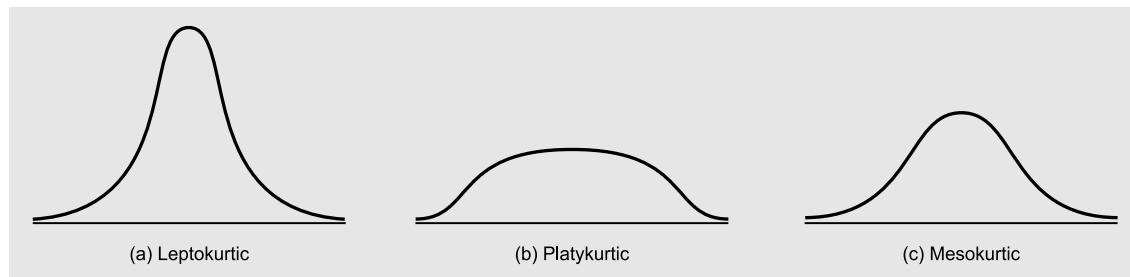
Kurtosis: The degree of flatness or peakedness in the region around the mode of a frequency curve.

The word ‘**kurtosis**’ comes from a Greek word meaning ‘humped’. In statistics, it refers to the degree of flatness or peakedness in the region about the mode of a frequency curve. A few definitions of kurtosis are as follows:

- *The degree of kurtosis of a distribution is measured relative to the peakedness of a normal curve.* —Simpson and Kafka
- *A measure of kurtosis indicates the degree to which a curve of a frequency distribution is peaked or flat-topped.* —Croxten and Cowden
- *Kurtosis refers to the degree of peakedness of hump of the distribution.* —C. H. Meyers

Two or more distributions may have identical average, variation, and skewness, but they may show different degrees of concentration of values of observations around the mode, and hence may show different degrees of peakedness of the hump of the distributions as shown in Fig. 14.2.

Figure 14.2
Shape of Three Different Curves
Introduced by Karl Pearson



14.4.1 Measures of Kurtosis

Leptokurtic: A frequency curve that is more peaked than the normal curve.

Platykurtic: A frequency curve that is flat-topped than the normal curve.

Mesokurtic: A frequency curve that is a normal (symmetrical) curve.

The fourth standardized moment α_4 (or β_2) is a measure of flatness or peakedness of a single humped distribution (also called *Kurtosis*). For a normal distribution $\alpha_4 = \beta_2 = 3$ so that $\gamma_2 = 0$ and hence any distribution having $\beta_2 > 3$ will be peaked more sharply than the normal curve known as *leptokurtic* (narrow) while if $\beta_2 < 3$, the distribution is termed as *platykurtic* (broad).

The value of β_2 is helpful in selecting an appropriate measure of central tendency and variation to describe a frequency distribution. For example, if $\beta_2 = 3$, mean is preferred; if $\beta_2 > 3$ (leptokurtic distribution), median is preferred; while for $\beta_2 < 3$ (platykurtic distribution), quartile range is suitable.

Remark: W. S. Gosset, explained different shapes of frequency curves as: Platykurtic curves, like the platypas, are squat with short tails; leptokurtic curves are high with long tails like the Kangaroos noted for leaping.

Example 14.10: The first four moments of a distribution about the value 5 of the variable are 2, 20, 40, and 50. Show that the mean is 7. Also find the other moments, β_1 and β_2 , and comment upon the nature of the distribution.

Solution: From the data of the problem, we have

$$\mu'_1 = 2, \mu'_2 = 20, \mu'_3 = 40, \mu'_4 = 50 \text{ and } A = 5$$

Now the moments about the arbitrary point 5 are calculated as follows:

$$\text{Mean, } \bar{x} = \mu'_1 + A = 2 + 5 = 7$$

$$\text{Variance, } \mu_2 = \mu'_2 - (\mu'_1)^2 = 20 - (2)^2 = 16$$

$$\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2(\mu'_1)^3 = 40 - 3(2)(20) + 2(2)^3 = -64$$

$$\begin{aligned}\mu_4' &= \mu_4 - 4\mu_1'\mu_3' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4 \\ &= 50 - 4(2)(40) + 6(20)(2)^2 - 3(2)^4 = 162\end{aligned}$$

The two constants, β_1 and β_2 , calculated from central moments are as follows:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-64)^2}{(16)^3} = \frac{4096}{4096} = 1$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{162}{(16)^2} = \frac{162}{256} = 0.63$$

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{-64}{(16)^{3/2}} = -1 (< 0), \text{ distribution is negatively skewed.}$$

$$\gamma_2 = \beta_2 - 3 = 0.63 - 3 = -2.37 (< 0), \text{ distribution is platykurtic.}$$

Example 14.11: Find the standard deviation and kurtosis of the following set of data pertaining to kilowatt hours (kwh) of electricity consumed by 100 persons in a city.

Consumption (in kwh)	:	0–10	10–20	20–30	30–40	40–50
Number of users	:	10	20	40	20	10

Solution: The calculations for standard deviation and kurtosis are shown in Table 14.6.

Table 14.6 Calculations of Standard Deviation and Kurtosis

Consumption (in kwh)	Number of Users (f)	Mid-Value (m)	$d = (m - A)/10$ $= (m - 25)/10$	$-fd$	fd^2
0–10	10	5	-2	-20	40
10–20	20	15	-1	-20	20
20–30	40	25	0	0	0
30–40	20	35	1	20	20
40–50	10	45	2	20	40
	100			0	120

$$\bar{x} = A + \frac{\sum fd}{N} \times h = 25 + \frac{0}{100} \times 10 = 25$$

Since $\bar{x} = 25$ is an integer value, therefore we may calculate moments about the actual mean

$$\mu_r = \frac{1}{n} \sum f(x - \bar{x})^r = \frac{1}{n} \sum f(m - \bar{x})^r$$

Let $d = \frac{m - \bar{x}}{h}$ or $(m - \bar{x}) = hd$. Therefore

$$\mu_r = h^r \frac{1}{n} \sum fd^r ; \quad h = \text{width of class intervals}$$

The calculations for moments are shown in Table 14.7.

Table 14.7 Calculations for Moments

Mid-value (m)	Frequency (f)	$d = \frac{m - 25}{10}$	fd	fd^2	fd^3	fd^4
5	10	-2	-20	40	-80	160
15	20	-1	-20	20	-20	20
25	40	0	0	0	0	0
35	20	1	20	20	20	20
45	10	2	20	40	80	160
	100		0	120	0	360

Moments about the origin A = 25 are:

$$\mu_1 = h \frac{1}{N} \sum fd = 10 \times \frac{1}{100} = 0$$

$$\mu_2 = h^2 \frac{1}{N} \sum fd^2 = (10)^2 \frac{1}{100} \times 120 = 120$$

$$\mu_3 = h^3 \frac{1}{N} \sum fd^3 = (10)^3 \frac{1}{100} \times 0 = 0$$

$$\mu_4 = h^4 \frac{1}{N} \sum fd^4 = (10)^4 \frac{1}{100} \times 360 = 36,000$$

$$S.D. (\sigma) = \sqrt{\mu_2} = \sqrt{120} = 10.95$$

Karl Pearson's measure of kurtosis is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{36,000}{(120)^2} = 2.5$$

and therefore $\gamma_2 = \beta_2 - 3 = 2.5 - 3 = -0.50$

Since $\beta_2 < 3$ (or $\gamma_2 < 0$), distribution curve is platykurtic.

Example 14.12: Calculate the value of γ_1 and γ_2 from the following data and interpret them.

Profit (Rs in lakh) : 10–20 20–30 30–40 40–50 50–60

Number of companies : 18 20 30 22 10

Comment on the skewness and kurtosis of the distribution. [Kumaon Univ., MBA, 1999]

Solution: Calculations for moments about an arbitrary constant value are shown in Table 14.8.

Table 14.8 Calculations of Moments

Profit (Rs lakh)	Mid-value (m)	Number of Companies (f)	$d = (m - 35)/10$	fd	fd^2	fd^3	fd^4
10–20	15	18	-2	-36	72	-144	288
20–30	25	20	-1	-20	20	-20	20
30–40	35	30	0	0	0	0	0
40–50	45	22	1	22	22	22	22
50–60	55	10	2	20	40	80	160
		100		-14	154	-62	490

$$\mu'_1 = \frac{\sum fd}{N} \times h = \frac{-14}{100} \times 10 = -1.4;$$

$$\mu'_2 = \frac{\sum fd^2}{N} \times h^2 = \frac{154}{100} \times 100 = 154$$

$$\mu'_3 = \frac{\sum fd^3}{N} \times h^3 = \frac{-62}{100} \times 1000 = -620;$$

$$\mu'_4 = \frac{\sum fd^4}{N} \times h^4 = \frac{490}{100} \times 10,000 = 49,000$$

The central moments are as follows:

$$\begin{aligned}\mu_2 &= \mu'_2 - (\mu'_1)^2 = 154 - (-1.4)^2 = 152.04 \\ \mu_3 &= \mu'_3 - 3\mu'_1\mu'_2 + 2(\mu'_1)^3 \\ &= -620 - 3(-1.4)(154) + 2(-1.4)^3 = 21.312 \\ \mu_4 &= \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'_2(\mu'_1)^2 - 3(\mu'_1)^3 \\ &= 49,000 - 4(-1.4)(-620) + 6(154)(-1.4)^2 - 3(-1.4)^3 = 47,327.51\end{aligned}$$

Karl Pearson's relative measure of skewness and kurtosis are as follows:

$$\begin{aligned}\text{Measure of skewness, } \gamma_1 &= \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{21.312}{(152.04)^{3/2}} = \frac{21.312}{1874.714} = 0.0114 \\ \text{Measure of kurtosis, } \beta_2 &= \frac{\mu_4}{\mu_2^2} = \frac{47,327.51}{(152.04)^2} = 2.047 \\ \gamma_2 &= \beta_2 - 3 = 2.047 - 3 = -0.953\end{aligned}$$

The value of $\gamma_1 = 0.0114$ suggests that the distribution is almost symmetrical and $\gamma_2 = -0.953 (< 0)$ indicates a platykurtic frequency curve.

Conceptual Questions 14B

10. What do you understand by the terms skewness and kurtosis? Point out their role in analysing a frequency distribution. [Delhi Univ., MBA, 1994]
11. Averages, dispersion, skewness, and kurtosis are complementary to one another in understanding a frequency distribution? Elucidate.
12. Explain how the measure of skewness and kurtosis can be used in describing a frequency distribution. [Delhi Univ., MBA, 1991]
13. Define moments. Establish the relationship between the moments about mean and moments about any arbitrary point.
14. Explain the terms 'skewness' and 'kurtosis' used in connection with the frequency distribution of a continuous variable. Give the different measures of skewness (any two of the measures to be given) and kurtosis.
15. What do you mean by 'kurtosis' in statistics? Explain one of the methods of measuring it.
16. What is meant by 'moments' of a frequency distribution? Show how moments are used to describe the characteristics of a distribution, that is, central tendency, dispersion, skewness, and kurtosis. [Delhi Univ., MBA, 1997]
17. How do measures of central tendency, dispersion, skewness, and kurtosis help in analysing a frequency distribution? Explain with the help of an example. [Sukhadia Univ., MBA, 1999]
18. In what way measures of central tendency, variation, skewness and kurtosis are complementary to one another in understanding a frequency distribution? Elucidate. [Osmania Univ., MBA, 1995]
19. A frequency distribution can be described almost completely by the first four moments and two measures based on moments. Examine.

Self-Practice Problems 14B

- 14.15 The first two moments of a distribution about the value 5 of the variable are 2 and 20. Find the mean and the variance.
 - 14.16 In a certain distribution the first four moments about the point 4 are 15, 17, -30, and 108 respectively. Find the kurtosis of the frequency curve and comment on its shape.
 - 14.17 Find the first four moments about the mean for the set of numbers 2, 4, 6, and 8.
 - 14.18 Explain whether the following results of a piece of computation for obtaining the second central moment are consistent or not; $n = 120$, $\Sigma fx = -125$, $\Sigma fx^2 = 128$.
 - 14.19 The first four central moments are 0, 4, 8, and 144. Examine the skewness and kurtosis.
 - 14.20 The central moments of a distribution are given by $\mu_2 = 140$, $\mu_3 = 148$, $\mu_4 = 6030$. Calculate the moment measures of skewness and kurtosis and comment on the shape of the distribution.
 - 14.21 Calculate β_1 and β_2 (measure of skewness and kurtosis) for the following frequency distribution and hence comment on the type of the frequency distribution:
- | | | | | | |
|-------|---|---|---|---|---|
| $x :$ | 2 | 3 | 4 | 5 | 6 |
| $f :$ | 1 | 3 | 7 | 2 | 1 |

- 14.22** Compute the first four moments about the mean from the following data:

Mid-value of variate : 5 10 15 20 25 30 35
Frequency : 8 15 20 32 23 17 5

Comment upon the nature of the distribution.

- 14.23** A record was kept over a period of 6 months by a sales manager to determine the average number of calls made per day by his six salesmen. The results are shown below:

Salesmen : A B C D E F
Average number of calls per day : 8 10 12 15 7 5

- (a) Compute a measure of skewness. Is the distribution symmetrical?
(b) Compute a measure of kurtosis. What does this measure mean?

- 14.24** Find the second, third, and fourth central moments of the frequency distribution given below. Hence find the measure of skewness and a measure of kurtosis of the following distribution:

Class limits	Frequency
100–104.9	7
105–109.9	13
110–114.9	25
115–119.9	25
120–124.9	30

- 14.25** Find the first four moments about the mean for the following distribution:

Class Interval : 60–62 63–65 66–68 69–71 72–74
Frequency : 5 18 42 27 8

- 14.26** Find the variance, skewness, and kurtosis of the following frequency distribution by the method of moments:

Class interval : 0–10 10–20 20–30 30–40
Frequency : 1 4 3 2

- 14.27** Find the kurtosis for the following distribution

Class interval : 0–10 10–20 20–30 30–40
Frequency : 1 3 4 2

Comment on the nature of the distribution.

Hints and Answers

- 14.15** Given $\mu'_1 = 2$, $\mu'_2 = 20$, $A = 5$; $\bar{x} = \mu'_1 + A = 7$; $\mu_2 (= \sigma^2) = \mu'_2 - (\mu'_1)^2 = 16$

- 14.16** $\mu'_1 = 1.5$, $\mu'_2 = 17$, $\mu'_3 = -30$ and $\mu'_4 = 108$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2(\mu'_1)^2 - 3(\mu'_1)^4}{\{\mu'_2 - (\mu'_1)^2\}^2} = 2.308;$$

distribution is platykurtic.

- 14.17** $\mu_1 = 0$, $\mu_2 = 5$, $\mu_3 = 0$ and $\mu_4 = 41$

- 14.18** $\mu_2 = \sum fx^2/N - (\sum fx/N)^2 = 128/120 - (-125/120)^2 = -0.0146$

since σ^2 cannot be negative, therefore the data is inconsistent.

- 14.19** $\beta_1 = \mu_3^2/\mu_2^3 = 2/(4)^3 = 1$; $\gamma_1 = +\sqrt{\beta_1} = 1$

$$\beta_2 = \mu_4/\mu_2^2 = 144/(4)^2 = 9; \quad \gamma_2 = \beta_2 - 3 = 6$$

- 14.20** $\beta_1 = \mu_3^2/\mu_2^3 = (148)^2/(140)^3$; $\gamma_1 = +\sqrt{\beta_1} = 0.089$

(Approximately symmetrical and platykurtic)

$$\beta_2 = \mu_4/\mu_2^2 = 6030/(140)^2 = 0.3076$$

- 14.21** Let $A = 4$; $\mu'_1 = \sum fd/\sum f = -0.07$

$$\mu'_2 = \sum f d^2 / \sum f = 0.92;$$

$$\mu'_3 = \sum f d^3 / \sum f = -0.07; \quad \mu'_4 = \sum f d^4 / \sum f = 2.64$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2 = 0.924;$$

$$\mu_3 = \mu'_3 - 2\mu'_2\mu'_1 + 2\mu'_1^3 = 0.123$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + \mu'_2\mu'_1^2 - 3\mu'_1^4 = 2.691$$

$$\beta_1 = \mu_3^2/\mu_2^3 = 0.019; \quad \beta_2 = \mu_4/\mu_2^2 = 3 \text{ (approx.)}$$

The distribution is approximately normal.

- 14.23** $\beta_1 = 0.11$; $\beta_2 = 1.97$

- 14.24** $\mu_2 = 54$; $\mu_3 = 100.5$, $\mu_4 = 7827$;

$$\gamma_1 = +\sqrt{\beta_1} = 0.2533; \quad \gamma_2 = \beta_2 - 3 = -0.3158$$

- 14.25** $\mu_1 = 0$, $\mu_2 = 8.527$, $\mu_3 = -2.693$, $\mu_4 = 199.375$

- 14.26** $\sigma^2 = \mu_2 = 84$, $\gamma_1 = +\sqrt{\beta_1} = 0.0935$; $\beta_2 = 2.102$

- 14.27** $\mu_2 = 81$, $\mu_3 = 14817$; $\beta_2 = \mu_4/\mu_2^2 = 2.26$

Formulae Used

1. Absolute measure of skewness

$$Sk = \bar{x} - \text{Mode} \text{ or } Q_3 + Q_1 - 2 \text{ Med}$$

2. Coefficient of skewness

Karl Pearson's

$$Sk_p = \frac{\bar{x} - Mo}{\sigma} \quad \text{or} \quad \frac{3(\bar{x} - \text{Med})}{\sigma}$$

$$\text{Bowley's, } Sk_b = \frac{Q_3 + Q_1 - 2\text{Med}}{Q_3 - Q_1}$$

$$\text{Kelly's, } Sk_k = \frac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}} \quad \text{or} \quad \frac{D_9 + D_1 - 2D_5}{D_9 - D_1}$$

3. Coefficient of skewness based on moments

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}; \quad \beta_2 = \frac{\mu_4}{\mu_2^2}$$

4. Moments

About the mean (origin)

$$\mu_r = \frac{1}{n} \sum (x - \bar{x})^r, r = 1, 2, 3, 4$$

About an arbitrary point, A

$$\mu'_r = \frac{1}{n} \sum (x - A)^r, r = 1, 2, 3, 4$$

$$5. \text{ Kurtosis} \quad \gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}}$$

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3$$

6. For a normal curve, $\beta_2 = 3$ or $\gamma_2 = 0$; for a leptokurtic curve, $\beta_2 > 3$ or $\gamma_2 > 0$ and for a platykurtic curve, $\beta_2 < 3$ or $\gamma_2 < 0$.

Review Self-Practice Problems

- 14.28** Calculate the first four moments about the mean and also the value of β_1 and β_2 from the following data:

Marks :

0–10 10–20 20–30 30–40 40–50 50–60 60–70

Number of students :

8 12 20 30 15 10 5

[Kumaon Univ., MBA, 1998]

- 14.29** From the following data calculate moments about (a) assumed mean, 25 (b) actual mean, and (c) moments about zero from the following data:

Variable : 0–10 10–20 20–30 30–40

Frequency : 1 3 4 2

[MD Univ., BCom, 1997]

- 14.30** The first four moments of a distribution about $x = 2$ are 1, 2.5, 5.5, and 16. Calculate the four moments about \bar{x} and about zero.

[Delhi Univ., MCom; MD Univ., MCom, 1999]

- 14.31** The first four central moments of distribution are 0, 2.5, 0.7, and 18.75. Comment on the skewness and kurtosis of the distribution. [Kanpur Univ., MCom, 1998]

- 14.32** Using moments, calculate a measure of relative skewness and a measure of relative kurtosis for the following distribution and comment on the result obtained:

Daily Wages (in Rs)	No. of Workers	Daily Wages (in Rs)	No. of Workers
70 but below 90	8	130 but below 150	9
90 but below 110	11	150 but below 170	4
110 but below 130	18		

[Kerala Univ., BCom, 1998]

- 14.33** Find the coefficient of skewness from the following information:

Difference of two quartiles = 8; Mode = 1;

Sum of two quartiles = 22; Mean = 8.

[Delhi Univ., BCom (H), 1997]

- 14.34** From the data given below calculate the coefficient of variation:

Karl Pearson's coefficient of skewness = 0.42

Arithmetic mean = 86

Median = 80

[Osmania Univ., BCom, 1998]

- 14.35** From the following data of the wages of 50 workers of a factory, compute the first four moments about mean and also the value of β_1 and β_2 . Comment on the results

Weekly Wages (Rs)	Number of Workers	Weekly Wages (Rs)	Number of Workers
100–120	1	180–200	12
120–140	3	200–220	4
140–160	7	220–240	3
160–180	20		

[Kurukshetra Univ., BCom, 1996]

- 14.36** In a frequency distribution, the coefficient of skewness based on quartiles is 0.6. If the sum of upper and lower quartiles is 100 and the median is 38, find the value of the upper quartile.

- 14.37** The following data are given to an economist for the purpose of economic analysis. The data refer to the length of a certain type of battery:

$$n = 100, \quad \sum fd = 50, \quad \sum fd^2 = 1970,$$

$$\sum fd^3 = 2948, \quad \sum fd^4 = 86,752$$

where $d = (x - 48)$. Do you think that the distribution is platykurtic? [Delhi Univ., B.Com (H) 1998]

- 14.38** The daily expenditure (in Rs) of 100 families is given below

Daily expenditure :

0–20 20–40 40–60 60–80 80–100

Number of families :

13 f_2 27 f_4 16

If mode of the distribution is 44, calculate Karl Pearson's coefficient of skewness.

- 14.39** Pearson's coefficient of skewness for a distribution is 0.4 and coefficient of variance is 30 per cent. Its mode is 88. Find the mean and median.

- 14.40** Calculate β_1 and β_2 from the frequency distribution and interpret the results.

Age	Frequency	Age	Frequency
25–30	2	45–50	25
30–35	8	50–55	16
35–40	18	55–60	7
40–45	27	60–65	2

[Kumaon Univ., MBA, 2003]

- 14.41** The following table gives the distribution of monthly wages of 500 workers in a factory:

Monthly Wages (Rs hundred)	Number of Workers	Monthly Wages (Rs hundred)	Number of Workers
15–20	10	30–35	220
20–25	25	35–40	70
25–30	145	40–45	30

Compute Karl Pearson's and Bowley's coefficient of skewness. Interpret your answer.

[Delhi Univ., MBA 2002]

- 14.42** The first two moments of a distribution about the value 5 of the variable are 2 and 20. Find the mean and variance.

Hints and Answers

14.28 $\mu'_1 = -1.8$, $\mu'_2 = 240$, $\mu'_3 = -1020$,
 $\mu'_4 = 1,44,000$
 $\mu_2 = 236.76$, $\mu_3 = 264.336$, $\mu_4 = 1,41,290.11$
 $\beta_1 = 0.005$ and $\beta_2 = 2.521$

14.29 (a) Moments about assumed mean, $A = 25$
 $\mu'_1 = -3$, $\mu'_2 = 90$, $\mu'_3 = -900$, $\mu'_4 = 21,000$.

(b) Moments about actual mean

$$\mu_1 = 0, \mu_2 = 81, \mu_3 = -144, \mu_4 = 14,817$$

(c) Moments about zero:

$$\begin{aligned} v_1 &= A + \mu'_1 = 25 - 3 = 22 \text{ (mean value)} \\ v_2 &= \mu_2 + (v_1)^2 = 565; \quad v_3 = \mu_3 + 3v_1^2 v_2 - 2v_1^3 \\ &= 15,850 \\ v_4 &= \mu_4 + 4v_1 v_3 - v_2 - 6v_1^2 + 3v_1^4 = 4,71,625 \end{aligned}$$

14.30 Given $\mu'_1 = 1$, $\mu'_2 = 2.5$, $\mu'_3 = 5.5$ and $\mu'_4 = 16$ and $A = 2$

Moments about mean:

$$\begin{aligned} \mu_2 &= \mu'_2 - (\mu'_1)^2 = 1.5; \\ \mu_3 &= \mu'_3 - 3\mu'_1 \mu'_2 + 3(\mu'_1)^2 = 0 \\ \mu_4 &= \mu'_4 - 4\mu'_1 \mu'_3 + 6(\mu'_1)^2 \mu'_2 - 3(\mu'_1)^4 = 6 \end{aligned}$$

Moments about zero:

$$\begin{aligned} v_1 &= A + \mu'_1 = 2 + 1 = 3; \quad v_2 = 10.5; \\ v_3 &= 40.5 \text{ and } v_4 = 168 \end{aligned}$$

14.31 Given $\mu_1 = 0$, $\mu_2 = 2.5$, $\mu_3 = 0.7$ and $\mu_4 = 18.75$

Measure of skewness, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0.031 (> 0)$, the distribution is slightly positively skewed.

Measure of kurtosis, $\beta_2 = \frac{\mu_4}{\mu_2^2} = 3$, the distribution is mesokurtic.

14.32 Moments about arbitrary point

$$\begin{aligned} \mu'_1 &= -2; \quad \mu'_2 = 136; \quad \mu'_3 = -680 \text{ and} \\ \mu'_4 &= 42,400 \end{aligned}$$

Moments about mean:

$$\mu_2 = 132, \mu_3 = 120 \text{ and } \mu_4 = 40,176$$

Measure of skewness, $\beta_1 = 0.006$ and measure of kurtosis, $\beta_2 = 2.306$

14.33 Mode = 3 Median – 2 Mean or $11 = 3 \text{ Med} - 2 \times 8$ or $\text{Med} = 9$

$$Q_3 + Q_1 = 22 \text{ and } Q_3 - Q_1 = 8, \text{ i.e., } Q_3 = 15, Q_1 = 7$$

$$\text{Coefficient of skewness} = \frac{Q_3 + Q_1 - 2\text{Med}}{Q_3 - Q_1}$$

$$= \frac{15 + 7 - 2(9)}{8} = 0.5$$

14.34 Mode = 3 Median – 2 Mean = $3(80) - 2(86) = 68$

$$\text{Coefficient of skewness} = \frac{\bar{x} - \text{Mode}}{\sigma}$$

$$\text{or } 0.42 = \frac{86 - 68}{\sigma} \text{ or } \sigma = 42.86$$

$$\begin{aligned} \text{Coefficient of variation (CV)} &= \frac{\sigma}{\bar{x}} \times 100 \\ &= \frac{42.86}{68} \times 100 \\ &= 49.84 \text{ per cent.} \end{aligned}$$

14.35 Moments about arbitrary mean

$$\mu'_1 = \frac{\sum fd}{n} \times h = 2.6; \quad \mu'_3 = \frac{\sum fd^3}{n} \times h^3 = 1340$$

$$\mu'_2 = \frac{\sum fd^2}{n} \times h^2 = 166;$$

$$\mu'_4 = \frac{\sum fd^4}{n} \times h^4 = 91,000.$$

Moments about mean

$$\begin{aligned} \mu_1 &= 0; \quad \mu_2 = 159.24; \quad \mu_3 = 80.352, \\ \mu_4 &= 83,659.87 \end{aligned}$$

$$\beta_1 = \mu_3^2 / \mu_2^3 = 0.0016$$

(distribution is almost symmetrical)

$\beta_2 = \mu_4 / \mu_2^2 = 3.3 (> 3)$, distribution is platykurtic.

14.36 Given $\text{Sk} = 0.6$, $Q_1 + Q_3 = 100$, $\text{Med} = 38$

$$\text{Sk}_b = \frac{Q_3 + Q_1 - 2\text{Med}}{Q_3 - Q_1} \text{ or } 0.6 = \frac{100 - 2 \times 38}{Q_3 - Q_1}$$

$$= \frac{100 - 76}{Q_3 - (100 - Q_3)} \text{ or } Q_3 = 70$$

14.37 $\mu_2 = \mu'_2 - (\mu'_1)^2 = \frac{\sum fd^2}{n} - \left[\frac{\sum fd}{n} \right]^2 = 19.7 - (0.5)^2 = 19.45$

$$\begin{aligned}\mu_4 &= \mu'_4 - 4\mu'_1\mu'_3 + 6(\mu'_1)^2\mu'_2 - 3(\mu'_1)^4 \\&= \frac{\sum fd^4}{n} - \frac{4 \sum fd}{n} \times \frac{\sum fd^3}{n} + \left(\frac{\sum fd}{n} \right)^2 \frac{\sum fd^2}{n} - 3 \left(\frac{\sum fd}{n} \right)^4 \\&= 867.52 - 4(0.5)(29.48) + 6(19.7)(0.5) - 3(0.5)^4 \\&= 837.92\end{aligned}$$

$$\therefore \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{837.92}{(19.45)^2} = 2.214 (< 3), \text{ distribution is platykurtic.}$$

14.38 Let the frequency for the class 20–40 be f_2 . Then frequency for the class 60–80 will be

$$f_4 = 100(13 + f_2 + 27 + 16) = 44 - f_2$$

Expenditure	Number of Families (f)	Cumulative Frequency (cf)
0–20	13	13
20–40	f_2	$13 - f_2$
40–60	27	$40 - f_2$
60–80	$44 - f_2$	84
80–100	16	100

$$\begin{aligned}\text{Mode} &= l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h \\&= 40 + \frac{27 - f_2}{54 - f_2 - 44 + f_2} \times 20 \text{ or } f_2 = 25\end{aligned}$$

Thus frequency for the class 20–40 is 25 and for the class 60–80 is $44 - 25 = 19$

$$\text{Apply the formula, } Sk_p = \frac{\bar{x} - Mo}{\sigma} = \frac{50 - 44}{25.3} = 0.237$$

14.39 Given $Sk_p = 0.4$, $CV = 0.30$, $Mode = 88$

$$Sk_p = \frac{\bar{x} - Mo}{\sigma} = \frac{1 - (M_0 / \bar{x})}{(\sigma / \bar{x})} = \frac{1 - (88 / \bar{x})}{0.30};$$

$$CV = \sigma / \bar{x} \text{ or } 0.30 = \sigma / \bar{x}$$

$$\frac{88}{\bar{x}} = 1 - 0.4 \times 0.3 = 0.88 \text{ or } \bar{x} = 100$$

Also, $Mode = 3 \text{ Med} - 2 \bar{x}$ or $88 = 3 \text{ Med} - 2(100)$ or $\text{Med} = 96$

$$\mathbf{14.40} \quad \beta_1 = \mu_3^2 / \mu_2^3 = (0.1955)^2 / (2.238)^3 = 0.0034;$$

$$\beta_2 = \mu_4 / \mu_2^2 = 12.966 / (2.238)^2 = 2.59$$

14.42 Given $A = 5$, $\mu'_1 = 2$, and $\mu'_2 = 20$.

$$\text{Mean} = A + \mu'_1 = 7 \text{ and variance, } \mu_2 = \mu'_2 - (\mu'_1)^2 = 16$$

This page is intentionally left blank.

People can be divided into three groups: those who make things happen, those who watch things happen, and those who wonder what happened.

—John W. Newbern

An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.

—John Tukey

Chi-Square and Other Non-Parametric Tests

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- recognize the advantage and disadvantages of non-parametric statistical tests
- learn how a non-parametric statistical test is conducted when
 - variables are measured on a nominal scale, and
 - measurements are of independent nature.
- test significant association between categorical variables.

15.1 INTRODUCTION

A variety of statistical tests are available for analysing a given set of data. An appropriate statistical test for analysing a given set of data is selected on the basis of

- scale of measurement of the data
- dependence/independence of the measurements
- number of populations being studied
- specific requirements such as sample size, shape of population distribution, and so on, for using a statistical test.

Generally, parametric and non-parametric statistical tests are distinguished on (i) the basis of the scaling of the data and (ii) the assumptions regarding the sampling distribution of sample statistic.

In Chapter 10, we discussed how z , t , and F test statistics are used for estimation and test of hypotheses about population parameters. The use of these tests

- (i) require the level of measurement attained on the collected data in the form of an interval scale or ratio scale,
- (ii) involve hypothesis testing of specified parameter values, and,
- (iii) require assumptions about the population distribution, in particular, assumption of normality and whether standard deviation of sampling/population distribution is known or not.

If these assumptions are not justified then these tests would not yield accurate conclusions about population parameters. In such circumstances, it is necessary to use few other hypothesis testing procedures that do not require these conditions to be met. These

procedures are referred to as *non-parametric tests*. **Non-parametric tests** (i) do not depend on the form of the underlying population distribution from which the samples were drawn, and (ii) use data that are of insufficient strength, i.e. data are categorical (nominally) scaled or ranks (ordinally) scaled.

A non-parametric procedure (or method), also called *distribution free test* satisfies at least one of the following criteria:

- (i) The procedure does not take into consideration any population parameter such as μ , σ or p .
- (ii) The procedure is applied only on categorical data that are non-numerical and frequency counts of categories for one or more variables.
- (iii) The procedure does not depend on the form of the underlying population distribution, in particular, the requirement of normality.

Non-parametric (distribution-free)

tests: The tests which can be used validly when the assumptions needed for parametric testing cannot be met.

15.2 ADVANTAGES AND LIMITATIONS OF NON-PARAMETRIC METHODS

Advantages Few advantages of using non-parametric methods are as under:

- (i) Non-parametric methods can be used to analyse categorical (nominal scaling) data, rank (ordinal scaling) data and interval (ratio scaling) data
- (ii) Non-parametric methods are generally easy to apply and quick to compute when sample size is small.
- (iii) Non-parametric methods require few assumptions but are very useful when the scale of measurement is weaker than required for parametric methods. Hence these methods are widely used and yield a more general, broad-based conclusions.
- (iv) Non-parametric methods provide an approximate solution to an exact problem whereas parametric methods provide an exact solution to an approximate problem.
- (v) Non-parametric methods provide solution to problems that do not require to make the assumption that a population is distributed normally or any specific shape.

Limitations: Few major limitations of non-parametric methods are as under.

- (i) Non-parametric methods should not be used when all the assumption of the parametric methods can be met. However they are equally powerful when assumptions are met, when assumptions are not met these may be more powerful.
- (ii) Non-parametric methods require more manual computational time when sample size gets larger.
- (iii) Table values for non-parametric statistics are not as readily available as of parametric methods.
- (iv) Non-parametric tests are usually not as widely used and not well known as parametric tests.

15.3 THE CHI-SQUARE DISTRIBUTION

The term non-parametric does not mean that the population distribution under study has no parameters. All populations have certain parameters which define their distribution. In this section we will discuss the **chi-square (χ^2) test** which belongs to non-parametric category of methods, to test a hypothesis. The symbol χ is the Greek letter 'chi'. The sampling distribution of χ^2 is called χ^2 -distribution. Like other hypothesis testing procedures, the calculated value of χ^2 -test statistic is compared with its critical (or table) value to know whether the null hypothesis is true. The decision of accepting a null hypothesis is based on how 'close' the sample results are to the expected results.

The probability density function of χ^2 -distribution is given by

$$y = y_0 (\chi^2)^{(v/2)-1} e^{-\chi^2/2} \quad (15-1)$$

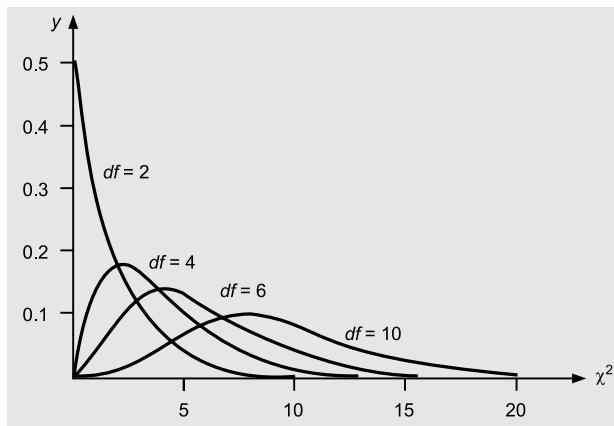
where v = degrees of freedom (dfv)

y_0 = a constant depending on degrees of freedom v

e = a constant, 2.71828

Chi-square test: A test for establishing the association between two categorical variables.

Figure 15.1
 χ^2 -Distributions with different values of df parameter.



The χ^2 -distribution is a continuous probability distribution extending from 0 to ∞ as shown in Fig. 15.1. Since χ^2 is the sum of squares, its value cannot be negative.

15.3.1 Properties of χ^2 Distribution

The following properties are useful while using χ^2 -test statistic to analyse its sampling distribution:

1. The shape of the curve for various values of degrees of freedom is shown in Fig 15.1. For $v = 1$, the density function (15-1) reduces to

$$y = y_0 e^{-\chi^2/2}$$

which is the standard normal curve for positive values of the variate.

2. The sampling distribution of χ^2 is a family of curves which vary with degrees of freedom. When $v = 1$, the curve is tangential to the x -axis at the origin, that is, the curve attains its maximum value when

$$\frac{dy}{d\chi^2} = y_0 [(v - 1) - \chi^2] \chi^{v-2} e^{-\chi^2/2} = 0$$

$$\text{or } (v - 1) - \chi^2 = 0 \quad \text{or} \quad \chi^2 = v - 1$$

when $v > 1$, the curve fall slowly and $y \rightarrow 0$ as $\chi^2 \rightarrow \infty$. In other words, sampling distribution of χ^2 is skewed towards higher values, that is, positively skewed.

3. For degrees of freedom $v = 3$, the curve touches the y -axis at the origin, and for all other values of $v > 4$, the curve is tangential to χ^2 axis at the origin.
4. For degrees of freedom $v \geq 30$, the χ^2 curve approximates to the normal curve with mean v and standard deviation $\sqrt{2v}$. In such a case the distribution of $\sqrt{2\chi^2}$ provides a better approximation to normality than χ^2 with mean $\sqrt{2v-1}$ and standard deviation one. This characteristic helps to test the significance of the difference between observed and expected values of the variable.
5. Since density function of χ^2 does not contain any parameter of population, χ^2 -test statistic is referred to as a non-parametric test. Thus χ^2 -distribution does not depend upon the form of the parent population.
6. The mean and variance of χ^2 -distribution are as follows:

$$\text{Mean, } \mu(\chi^2) = v \text{ and Variance, } \sigma^2(\chi^2) = 2v.$$

15.3.2 Conditions for the Applications of χ^2 Test

Before using χ^2 as a test statistic to test a hypothesis, the following conditions are necessary:

1. The experiment consists of n identical but independent trials. The outcome of each trial falls into one of k categories. The observed number of outcome in each category, written as O_1, O_2, \dots, O_n , with $O_1 + O_2 + \dots + O_n = 1$ are counted.
2. If there are only two cells, the expected frequency in each cell should be 5 or more. Because for observations less than 5, the value of χ^2 shall be over estimated, resulting in the rejection of the null hypothesis.

3. For more than two cells, if more than 20 per cent of the cells have expected frequencies less than 5, then χ^2 should not be applied.
4. Samples must be drawn randomly from the population of interest. All the individual observations in a sample should be independent.
5. The sample should contain at least 50 observations.
6. The data should be expressed in original units, rather than in percentage or ratio form. Such precaution helps in comparison of attributes of interest.

15.4 THE CHI-SQUARE TEST-STATISTIC

Like t and F distributions, a χ^2 -distribution is also a function of its degrees of freedom. This distribution is skewed to the right and the random variable can never take a negative value. Theoretically, its range is from 0 to ∞ as shown in Fig. 15.2. Values of χ^2 that divide the curve with a proportion of the area equivalent to α (level of significance) in the right tail are given in the Appendix. The χ^2 -test statistic is given by

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad (15-2)$$

where O = an observed frequency in a particular category
 E = an expected frequency for a particular category

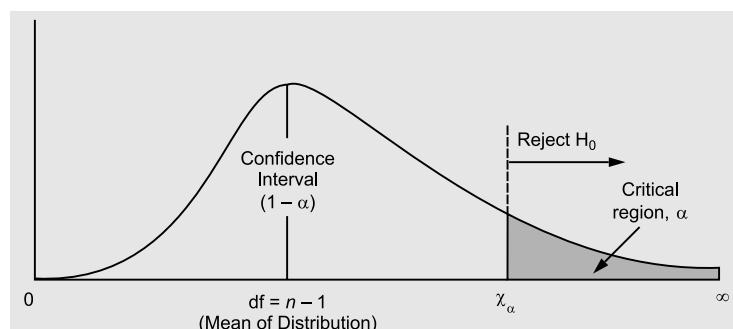


Figure 15.2
 χ^2 Density Function

Decision Rule The calculated value of χ^2 is compared with its critical value at a particular level of significance and degrees of freedom. If $\chi_{\text{cal}}^2 > \chi_{\text{critical}}^2$, then the null hypothesis is rejected in favour of the alternative hypothesis, and it is concluded that the difference between two sets of frequencies is significant.

The degrees of freedom for χ^2 -test statistic depend on the test and certain other factors, which will be discussed later in this chapter.

Since the mean of χ^2 -distribution is equal to the number of degrees of freedom, therefore skewness of this distribution is considerable when the number of degrees of freedom is small, but it reduces as the number of degrees of freedom increases as shown in Fig. 15.2.

15.4.1 Grouping of Small Frequencies

One or more observations with frequencies less than 5 may be grouped together to represent a single category before calculating the difference between observed and expected frequencies. For example, the figures given below are the theoretical (observed) and expected frequencies (based on Poisson distribution) having same mean value and equal number of total frequencies.

Observed frequencies :	305	365	210	80	28	$\overline{\begin{array}{cc} 9 & 3 \end{array}}$
Expected frequencies :	301	361	217	88	26	$\overline{\begin{array}{cc} 6 & 1 \end{array}}$

These 7 classes can be reduced to 6 by combining the last two frequencies in both the cases as follows:

Observed frequencies :	305	365	210	80	28	12
Expected frequencies :	301	361	217	88	26	7

Since the original 7 classes have been reduced to 6 by grouping, therefore the revised degrees of freedom are $df = 6 - 2 = 4$, due to two restraints.

15.5 APPLICATIONS OF χ^2 TEST

A few important applications of χ^2 test discussed in this chapter are as follows:

- Test of independence
- Test of goodness-of-fit
- Yate's correction for continuity
- Test for population variance
- Test for homogeneity

15.5.1 Contingency Table Analysis : Chi-Square Test of Independence

The χ^2 test of independence is used to analyse the frequencies of two qualitative variables or attributes with multiple categories to determine whether the two variables are independent. The chi-square test of independence can be used to analyse any level of measurement, but it is particularly useful in analysing nominal data. For example,

- Whether voters can be classified by gender is independent of the political affiliation
- Whether university students classified by gender are independent of courses of study
- Whether wage-earners classified by education level are independent of income
- Whether type of soft drink preferred by a consumer is independent of the consumer's age.
- Whether absenteeism is independent of job classification
- Whether an item manufactured is acceptable or not is independent of the shifts in which it was manufactured.

When observations are classified according to two qualitatives variables or attributes and arranged in a table, the display is called a **contingency table** as shown in Table 15.1. The test of independence uses the contingency table format and is also referred to as a *Contingency Table Analysis (or Test)*.

Table 15.1: Contingency Table

Variable B	Variable A				Total
	A_1	A_2	\dots	A_c	
B_1	O_{11}	O_{12}	\dots	O_{1c}	R_1
B_2	O_{21}	O_{22}	\dots	O_{2c}	R_2
.	.	.			.
.	.	.			.
B_r	O_{r1}	O_{r2}	\dots	O_{rc}	R_r
Total	C_1	C_2	\dots	C_c	N

It may be noted that the variables A and B have been classified into mutually exclusive categories. The value O_{ij} is the observed frequency for the cell in row i and colum j . The row and column totals are the sums of the frequencies. The row and column totals are added up to get a grand total n , which represents the sample size.

The *expected frequency*, E_{ij} , corresponding to an observed frequency O_{ij} in row i and column j under the assumption of independence, is based on the multiplicative rule of probability. That is, if two events A_i and B_j are independent, then the probability of their joint occurrence is equal to the product of their individual probabilities. Thus the expected frequencies in each cell of the contingency table are calculated as follows:

$$\begin{aligned}
 E_{ij} &= \frac{\text{Row } i \text{ total}}{\text{Sample size}} \times \frac{\text{Column } j \text{ total}}{\text{Sample size}} \times \text{Grand total} \\
 &= \frac{R_i}{N} \times \frac{C_j}{N} \times N = \frac{R_i \times C_j}{N}
 \end{aligned} \tag{15-3}$$

The analysis of a two-way contingency table helps to answer the question whether the two variables are unrelated or independent of each other. Consequently, *the null hypothesis for a chi-square test of independence is that the two variables are independent*. If null hypothesis H_0 is rejected, then two variables are not independent but are related. Hence, the χ^2 -test statistic measures how much the observed frequencies differ from the expected frequencies when the variables are independent.

The Procedure The procedure to test the association between two independent variables where the sample data is presented in the form of a contingency table with r rows and c columns is summarized as follows:

Step 1: State the null and alternative hypotheses

H_0 : No relationship or association exists between two variables, that is, they are independent

H_1 : A relationship exists, that is, they are related

Step 2: Select a random sample and record the observed frequencies (O values) in each cell of the contingency table and calculate the row, column, and grand totals.

Step 3: Calculate the expected frequencies (E -values) for each cell:

$$E = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

Step 4: Compute the value of test-statistic

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Step 5: Calculate the degrees of freedom. The degrees of freedom for the chi-square test of independence are given by the formula

$$df = (\text{Number of rows} - 1)(\text{Number of columns} - 1) = (r - 1)(c - 1)$$

Step 6: Using a level of significance α and df , find the critical (table) value of χ^2_α (see Appendix). This value of χ^2_α corresponds to an area in the right tail of the distribution.

Step 7: Compare the calculated and table values of χ^2 . Decide whether the variables are independent or not, using the decision rule:

- Accept H_0 if χ^2_{cal} is less than its table value $\chi^2_{\alpha, (r-1)(c-1)}$
- Otherwise reject H_0

Example 15.1: Two hundred randomly selected adults were asked whether TV shows as a whole are primarily entertaining, educational, or a waste of time (only one answer could be chosen). The respondents were categorized by gender. Their responses are given in the following table:

Gender	Opinion			Total
	Entertaining	Educational	Waste of time	
Female	52	28	30	110
Male	28	12	50	90
Total	80	40	80	200

Is this evidence convincing that there is a relationship between gender and opinion in the population interest?

Solution: Let us assume the null hypothesis that the opinion of adults is independent of gender.

The contingency table is of size 2×3 , the degrees of freedom would be $(2-1)(3-1) = 2$, that is, we will have to calculate only two expected frequencies and the other four can be automatically determined as shown below:

$$E_{11} = \frac{\text{Row 1 total} \times \text{Column 1 total}}{\text{Grand total}} = \frac{110 \times 80}{200} = 44$$

$$E_{12} = \frac{\text{Row 1 total} \times \text{Column 2 total}}{\text{Grand total}} = \frac{110 \times 40}{200} = 22$$

$$E_{13} = 110 - (44 + 22) = 44$$

$$E_{21} = 80 - E_{11} = 80 - 44 = 36$$

$$E_{22} = 40 - E_{12} = 40 - 22 = 18$$

$$E_{23} = 80 - E_{13} = 80 - 44 = 36$$

The contingency table of expected frequencies is as follows:

Gender	Opinion			Total
	Entertaining	Educational	Waste of time	
Male	44	22	44	110
Female	36	18	36	90
Total	80	40	80	200

Arranging the observed and expected frequencies in the following table to calculate the value of χ^2 -test statistic:

Observed (O)	Expected (E)	$(O - E)$	$(O - E)^2$	$(O - E)^2/E$
52	44	8	64	1.454
28	22	6	36	1.636
30	44	14	196	4.455
28	36	-8	64	1.777
12	18	-6	36	2.000
50	36	14	196	5.444
				16.766

The critical (or table) value of $\chi^2 = 5.99$ at $\alpha = 0.05$ and $df = 2$. Since the calculated value of $\chi^2 = 16.766$ is more than its critical value, the null hypothesis is rejected. Hence, we conclude that the opinion of adults is not independent of gender.

Example 15.2: A company is interested in determining whether an association exists between the commuting time of their employees and the level of stress-related problems observed on the job. A study of 116 assembly-line workers reveals the following:

Commuting Time	Stress			Total
	High	Moderate	Low	
Under 20 min	9	5	18	32
20 – 50 min	17	8	28	53
over 50 min	18	6	7	31
Total	44	19	53	116

At $\alpha = 0.01$ level of significance, is there any evidence of a significant relationship between commuting time and stress?

Solution: Let us assume the null hypothesis that stress on the job is independent of commuting time.

The contingency table is of size 3×3 , the degrees of freedom would be $(3 - 1)(3 - 1) = 4$, that is, we will have to calculate only four expected frequencies and the others can be calculated automatically as shown below:

$$E_{11} = \frac{32 \times 44}{116} = 12.14 \quad E_{12} = \frac{32 \times 19}{116} = 5.24 \quad E_{13} = 14.62$$

$$E_{21} = \frac{53 \times 44}{116} = 20.10 \quad E_{22} = \frac{53 \times 19}{116} = 8.68 \quad E_{23} = 24.22$$

$$E_{31} = \frac{31 \times 44}{116} = 11.75 \quad E_{32} = \frac{31 \times 19}{116} = 5.08 \quad E_{33} = 14.17$$

Arranging the observed and expected frequencies in the following table to calculate the value of χ^2 -test statistic:

<i>Observed (O)</i>	<i>Expected (E)</i>	<i>O - E</i>	<i>(O - E)²</i>	<i>(O - E)²/E</i>
9	12.14	-3.14	9.85	0.811
5	5.24	-0.24	0.05	0.009
18	14.62	3.38	11.42	0.781
17	20.10	-3.10	9.61	0.478
8	8.68	-0.68	0.45	0.052
28	24.22	3.78	14.28	0.589
18	11.75	6.25	39.06	3.324
6	5.08	0.92	0.84	0.165
7	14.17	-7.17	51.40	3.627
				9.836

The critical value of $\chi^2 = 13.30$ at $\alpha = 0.01$ and $df = 4$. Since calculated value of $\chi^2 = 9.836$ is less than its critical value, the null hypothesis H_0 is accepted. Hence we conclude that stress on the job is independent of commuting time.

Example 15.3: A certain drug is claimed to be effective in curing colds. In an experiment on 500 persons with cold, half of them were given the drug and half of them were given sugar pills. The patients' reactions to the treatment are recorded in the following table:

<i>Treatment</i>	<i>Consequence</i>			<i>Total</i>
	<i>Helped</i>	<i>Reaction</i>	<i>No effect</i>	
Drug	150	30	70	250
Sugar pills	130	40	80	250
Total	280	70	150	500

On the basis of the data, can it be concluded that there is a significant difference in the effect of the drug and sugar pills?

[Lucknow Univ., MBA, 1998, Delhi Univ., MBA, 1999, 2002]

Solution: Let us assume the null hypothesis that there is no significant difference in the effect of the drug and sugar pills.

The contingency table is of size 2×3 , the degrees of freedom would be $(2 - 1)(3 - 1) = 2$, that is, we would have to calculate only two expected frequencies and others can be automatically determined as shown below:

$$E_{11} = \frac{250 \times 280}{500} = 140; \quad E_{12} = \frac{250 \times 70}{500} = 35$$

The contingency table of expected frequencies is as follows:

<i>Treatment</i>	<i>Consequence</i>			<i>Total</i>
	<i>Helped</i>	<i>Harmed</i>	<i>No Effect</i>	
Drug	140	35	75	250
Sugar pills	140	35	75	250
Total	280	70	150	500

Arranging the observed and expected frequencies in the following table to calculate the value of χ^2 -test statistic.

<i>Observed (O)</i>	<i>Expected (E)</i>	$(O - E)$	$(O - E)^2$	$(O - E)^2/E$
150	140	10	100	0.714
130	140	-10	100	0.714
30	35	-5	25	0.714
40	35	5	25	0.714
70	75	-5	25	0.333
80	75	5	25	0.333
				3.522

The critical value of $\chi^2 = 5.99$ at $\alpha = 0.05$ and $df = 2$. Since the calculated value of $\chi^2 = 3.522$ is less than its critical value, the null hypothesis is accepted. Hence we conclude that there is no significant difference in the effect of the drug and sugar pills.

Self-Practice Problems 15A

- 15.1** In an anti-malaria campaign in a certain area, quinine was administered to 812 persons out of a total population of 3248. The number of fever cases reported is shown below:

<i>Treatment</i>	<i>Fever</i>	<i>No Fever</i>	<i>Total</i>
Quinine	20	792	812
No quinine	220	2216	2436
Total	240	3008	3248

Discuss the usefulness of quinine in checking malaria.

[MD Univ., MCom, 1998]

- 15.2** Based on information from 1000 randomly selected fields about the tenancy status of the cultivation of these fields and use of fertilizers, collected in an agro-economic survey, the following classifications were noted:

	<i>Owned</i>	<i>Rented</i>	<i>Total</i>
Using fertilizers	416	184	600
Not using fertilizers	64	336	400
Total	480	520	1000

Would you conclude that owner cultivators are more inclined towards the use of fertilizers at $\alpha = 0.05$ level of significance? Carry out the chi-square test as per testing procedures.

[Osmania Univ., MCom, 1997]

- 15.3** In an experiment on immunization of cattle from tuberculosis, the following results were obtained:

	<i>Affected</i>	<i>Not Affected</i>
Inoculated	12	26
Not inoculated	16	6

Calculate the χ^2 and discuss the effect of vaccine in controlling susceptibility to tuberculosis.

[Rajasthan Univ., MCom, 1996]

- 15.4** From the data given below about the treatment of 250 patients suffering from a disease, state whether the new treatment is superior to the conventional treatment.

<i>Treatment</i>	<i>No. of Patients</i>		
	<i>Favourable</i>	<i>Not Favourable</i>	<i>Total</i>
New	140	30	170
Conventional	60	20	80
Total	200	50	250

[MD Univ., MCom, 1998]

- 15.5** 1000 students at college level are graded according to their IQ and their economic conditions. Use chi-square test to find out whether there is any association between economic conditions and the level of IQ.

<i>Economic Conditions</i>	<i>IQ Level</i>			<i>Total</i>
	<i>High</i>	<i>Medium</i>	<i>Low</i>	
Rich	160	300	140	600
Poor	140	100	160	400
Total	300	400	300	1000

[Madurai Univ., MCom, 1996]

- 15.6** 200 digits are chosen at random from a set of tables. The frequencies of the digits are as follows:

Digit : 0 1 2 3 4 5 6 7 8 9

Frequency: 18 19 23 21 16 25 22 20 21 15

Use the χ^2 test to assess the correctness of the hypothesis that the digits were distributed in equal numbers in the tables from which they were chosen.

[HP Univ., MCom, 2000]

- 15.7** In an experiment on pea-breeding, Mendel obtained the following frequencies of seeds: 315 round and yellow, 101 wrinkled and yellow, 108 round and green, 32 wrinkled and green. According to his theory of heredity

the numbers should be in proportion 9 : 3 : 3 : 1. Is there any evidence to doubt the theory at $\alpha = 0.05$ level of significance?

[Delhi Univ., MCom, 1996]

- 15.8** Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different intelligence levels. The results are as follows:

Researcher	Below	Average	Above	Genius	Total
	Average		Average		
X	86	60	44	10	200
Y	40	33	25	2	100
Total	126	93	69	12	300

Would you say that the sampling techniques adopted by the two researchers are significantly different?

[Delhi Univ., MBA, 1998]

- 15.9** From the following data, find out whether there is any relationship between sex and preference of colour:

Colour	Males	Females	Total
Red	10	40	50
White	70	30	100
Green	30	20	50
Total	110	90	200

[HP Univ., MBA., 1998; Madurai Univ., MCom, 1999]

- 15.10** A manufacturer of TV sets was trying to find out what variables influenced the purchase of a TV set. Level of income was suggested as a possible variable influencing the purchase of TV sets. A sample of 500 households was selected and the information obtained is classified as shown below:

Income Group	Have TV Set	Do not have TV Set
Low income group	0	250
Middle income group	50	100
High income group	80	20

Is there evidence from the above data of a relation ownership of TV sets and level of income?

[Delhi Univ., MBA, 2000]

- 15.11** A marketing agency gives the following information about the age groups of the sample informants and their liking for a particular model of scooter which a company plans to introduce:

Choice	Age Group of Informants			Total
	Below 20	20–39	40–59	
Liked	125	420	60	605
Disliked	75	220	100	395
Total	200	640	160	1000

On the basis of above data, can it be concluded that the model appeal is independent of the age group of the informants?

[HP Univ., MBA, 1998]

Hints and Answers

- 15.1** Let H_0 : Quinine is not effective in checking malaria. $\chi^2_{\text{cal}} = 38.393$ is more than its critical value $\chi^2_{\text{critical}} = 3.84$ at $\alpha = 0.05$ and $df = 1$, the null hypothesis is rejected.
- 15.2** Let H_0 : Ownership of fields and the use of fertilizers are independent attributes. $\chi^2_{\text{cal}} = 273.504$ is more than its critical value $\chi^2_{\text{critical}} = 3.84$ at $\alpha = 0.05$ and $df = 1$, the null hypothesis is rejected.
- 15.3** Let H_0 : Vaccine is not effective in controlling susceptibility to tuberculosis. $\chi^2_{\text{cal}} = 7.796$ is more than its critical value $\chi^2_{\text{critical}} = 3.84$ at $\alpha = 0.05$ and $df = 1$, the null hypothesis is rejected.
- 15.4** Let H_0 : No significant difference between the new and conventional treatment. $\chi^2_{\text{cal}} = 1.839$ is less than its critical value $\chi^2_{\text{critical}} = 3.84$ at $\alpha = 0.05$ and $df = 1$, the null hypothesis is accepted.
- 15.5** Let H_0 : No association between economic conditions and the level of IQ. $\chi^2_{\text{cal}} = 65.277$ is more than its critical value $\chi^2_{\text{critical}} = 5.99$ at $\alpha = 0.05$ and $df = 2$, the null hypothesis is rejected.

- 15.6** Let H_0 : Digits were distributed in equal numbers. On the basis of H_0 , $E = 200/10 = 20$ as the frequency of occurrence for 0, 1, 2, ..., digits.

$\chi^2_{\text{cal}} = 4.3$ is less than its critical value $\chi^2 = 16.22$ at $\alpha = 0.05$ and $df = 10 - 1 = 9$, the null hypothesis is accepted.

- 15.7** Let H_0 : No significant difference in the observed and expected frequencies.

Calculations of expected values are as follows:

$$E_1 = \frac{556 \times 9}{16} = 312.75, E_2 = \frac{556 \times 3}{16} = 104.25,$$

$$E_3 = 104.25 \text{ and } E_4 = 34.75, \text{ respectively.}$$

Category	O	E	$(O-E)^2$	$(O-E)^2/E$
Round and yellow	315	312.75	5.062	0.016
Wrinkled and yellow	101	104.25	10.562	0.101
Round and green	108	104.25	14.062	0.135
Wrinkled and green	32	34.75	7.562	0.218
				0.470

Since $\chi^2_{\text{cal}} = 0.470$ is less than its critical value $\chi^2 = 7.82$ at $\alpha = 0.05$ and $df = 4 - 1 = 3$, the null hypothesis is accepted.

- 15.8** Let H_0 : Sampling techniques adopted by two researchers are not significantly different.

$\chi^2_{\text{cal}} = 2.098$ is less than its critical value $\chi^2_{\text{critical}} = 7.82$ at $\alpha = 0.05$ and $df = 3$, the null hypothesis is accepted.

- 15.9** Let H_0 : No relationship between gender and preference of colour.

$\chi^2_{\text{cal}} = 34.35$ is more than its critical value $\chi^2_{\text{critical}} = 5.99$ at $\alpha = 0.05$ and $df = 2$, the null hypothesis is rejected.

- 15.10** Let H_0 : The ownership of TV sets is independent of the level of income.

$\chi^2_{\text{cal}} = 243.59$ is more than its critical value $\chi^2_{\text{critical}} = 7.82$ at $\alpha = 0.05$ and $df = 2$, the null hypothesis is rejected.

- 15.11** Let H_0 : The model appeal is independent of the age group of the informants.

$\chi^2_{\text{cal}} = 42.788$ is more than its critical value $\chi^2_{\text{critical}} = 5.99$ at $\alpha = 0.05$ and $df = 2$, the null hypothesis is rejected.

15.5.2 Chi-Square Test for Goodness-of-Fit

Goodness-of-fit: A statistical test conducted to determine how closely the observed frequencies fit those predicted by a hypothesized probability distribution for population.

On several occasions a decision-maker needs to understand whether an actual sample distribution matches or coincides with a known theoretical probability distribution such as binomial, Poisson, normal, and so on. *The χ^2 test for goodness-of-fit is a statistical test of how well given data support an assumption about the distribution of a population or random variable of interest. The test determines how well an assumed distribution fits the given data.* To apply this test, a particular theoretical distribution is first hypothesized for a given population and then the test is carried out to determine whether or not the sample data could have come from the population of interest with the hypothesized theoretical distribution. The observed frequencies or values come from the sample and the expected frequencies or values come from the theoretical hypothesized probability distribution. The goodness-of-fit test now focuses on the differences between the observed values and the expected values. Large differences between the two distributions throw doubt on the assumption that the hypothesized theoretical distribution is correct. On the other hand, small differences between the two distribution may be assumed to be resulting from sampling error.

The Procedure The general steps to conduct a goodness-of-fit test for any hypothesized population distribution are summarized as follows:

Step 1: State the null and alternative hypotheses

H_0 : No difference between the observed and expected sets of frequencies.

H_1 : There is a difference

Step 2: Select a random sample and record the observed frequencies (O values) for each category.

Step 3: Calculate expected frequencies (E values) in each category by multiplying the category probability by the sample size.

Step 4: Compute the value of test statistic

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Step 5: Using a level of significance α and $df = n - 1$ provided that the number of expected frequencies are 5 or more for all categories, find the critical (table) value of χ^2 (See Appendix).

Step 6: Compare the calculated and table value of χ^2 , and use the following decision rule:

- Accept H_0 if χ^2_{cal} is less than its critical value $\chi^2_{\alpha, n-1}$
- Otherwise reject H_0

Example 15.4: A Personnel Manager is interested in trying to determine whether absenteeism is greater on one day of the week than on another. His records for the past year show the following sample distribution:

Day of the week : Monday Tuesday Wednesday Thursday Friday

No. of absentees : 66 56 54 48 75

Test whether the absence is uniformly distributed over the week.

[Madras Univ., MCom, 1996]

Solution: Let us assume the null hypothesis that the absence is uniformly distributed over the week.

The number of absentees during a week are 300 and if absenteeism is equally probable on all days, then we should expect $300/5 = 60$ absentees on each day of the week. Now arranging the data as follows:

Category	O	E	O - E	$(O - E)^2$	$(O - E)^2/E$
Monday	66	60	6	36	0.60
Tuesday	57	60	-3	9	0.15
Wednesday	54	60	-6	36	0.60
Thursday	48	60	-12	144	2.40
Friday	75	60	-15	225	3.75
					7.50

The critical value of $\chi^2 = 9.49$ at $\alpha = 0.05$ and $df = 5 - 1 = 4$. Since calculated value $\chi_{\text{cal}} = 7.50$ is less than its critical value, the null hypothesis is accepted.

Example 15.5: A survey of 800 families with 4 children each revealed following distribution:

No. of boys	:	0	1	2	3	4
No. of girls	:	4	3	2	1	0
No. of families	:	32	178	290	236	64

Is this result consistent with the hypothesis that male and female births are equally probable?

Solution: Let us assume the null hypothesis that male and female births are equally probable.

The probability of a male or female is $p = q = 1/2$. Since birth of male and female is mutually exclusive and exhaustive, the expected number of families having different combinations of boys and girls can be calculated using binomial probability distribution as follows:

$$\begin{aligned} P(x = r) &= {}^4C_r p^r q^{4-r}; r = 0, 1, \dots, 4 \\ &= {}^4C_r (1/2)^4 \text{ since } p = q = 1/2 \end{aligned}$$

The calculations for expected frequencies for each combination (category) of boy or girl are as shown below:

Category x	$P(x = r)$	Expected Frequency, $n P(x)$
0	${}^4C_0 (1/2)^4 = 1/16$	$800 \times (1/16) = 50$
1	${}^4C_1 (1/2)^4 = 4/16$	$800 \times (4/16) = 200$
2	${}^4C_2 (1/2)^4 = 6/16$	$800 \times (6/16) = 300$
3	${}^4C_3 (1/2)^4 = 4/16$	$800 \times (4/16) = 200$
4	${}^4C_4 (1/2)^4 = 1/16$	$800 \times (1/16) = 50$

To apply the χ^2 -test, arrange the observed and expected frequencies in the table below:

Category	O	E	O - E	$(O - E)^2$	$(O - E)^2/E$
0	32	50	-18	324	6.480
1	178	200	-22	484	2.420
2	290	300	-10	100	0.333
3	236	200	36	1296	6.480
4	64	50	14	196	3.920
					19.633

The critical value of $\chi^2 = 9.488$ at $\alpha = 0.05$ and $df = 5 - 1 = 4$. Since calculated value of χ^2 is greater than its table value, the hypothesis is rejected. Hence, we conclude that male and female births do not seem to be equally probable.

Example 15.6: The figures given below are (a) the theoretical frequencies of a distribution, and (b) the frequencies of the normal distribution having the same mean, standard deviation, and the total frequency as in (a):

(a)	1	5	20	28	42	22	15	5	2
(b)	1	6	18	25	40	25	18	6	1

Do you think that the normal distribution provides a good fit to the data?

Solution: Let us assume the null hypothesis that there is no difference between observed frequencies and expected frequencies obtained by normal distribution.

Since the observed and expected frequencies are less than 10 in the beginning and end of the series, we shall group these classes together as follows:

O	E	O - E	$(O - E)^2$	$(O - E)^2/E$
$\frac{1}{5} \left[= 6 \right]$	$\frac{1}{6} \left[= 7 \right]$	-1	1	0.143
20	18	2	4	0.222
28	25	3	9	0.360
42	40	2	4	0.100
22	25	-3	9	0.360
15	18	-3	9	0.500
$\frac{5}{2} \left[= 7 \right]$	$\frac{6}{1} \left[= 7 \right]$	0	0	0.000
				1.685

The revised degrees of freedom are $df = 9 - 1 - 4 = 4$ is 4. The critical value $\chi^2_{\text{critical}} = 9.49$ at $\alpha = 0.05$ and $df = 4$. Since the calculated value of $\chi^2 = 1.685$ is less than its critical value, the null hypothesis is accepted. Hence, we conclude that the normal distribution provides a good fit to the data.

Example 15.7: A book has 700 pages. The number of pages with various numbers of misprints is recorded below:

No. of misprints	:	0	1	2	3	4	5
No. of pages with misprints	:	616	70	10	2	1	1

Can a Poisson distribution be fitted to this data? [Delhi Univ., MCom, 1999]

Solution: Let us assume the null hypothesis that the data is fitted with Poisson distribution.

The number of pages in the book are 700, whereas the maximum possible number of misprints are only 5. Thus we may apply Poisson probability distribution to calculate the expected number of misprints in each page of the book as follows:

Mistakes (x)	Number of Pages (f)	fx
0	616	0
1	70	70
2	10	20
3	2	6
4	1	4
5	1	5
$n = 700$		105

$$\lambda \text{ or } m = \frac{\Sigma fx}{n} = \frac{105}{700} = 0.15$$

The calculations for expected frequencies for misprints from 0 to 5 are shown below:

$$\begin{aligned}
 P(x=0) &= e^{-\lambda} = e^{-0.15} = 0.8607 \text{ (from table)} \\
 nP(x=0) &= 700 \times 0.8607 = 602.5 \\
 nP(x=1) &= nP(x=0) \lambda = 602.5 \times 0.15 = 90.38 \\
 nP(x=2) &= nP(x=1) \frac{\lambda}{2} = 90.38 \times \frac{0.15}{2} = 6.78 \\
 nP(x=3) &= nP(x=2) \frac{\lambda}{3} = 6.78 \times \frac{0.15}{3} = 0.34 \\
 nP(x=4) &= nP(x=3) \frac{\lambda}{4} = 0.34 \times \frac{0.15}{4} = 0.013 \\
 nP(x=5) &= nP(x=4) \frac{\lambda}{5} = 0.013 \times \frac{0.15}{5} = 0
 \end{aligned}$$

To apply the χ^2 -test, arrange the observed and expected frequencies in the table below:

Mistake	O	E	O - E	$(O - E)^2$	$(O - E)^2/E$
0	616	602.5	13.50	182.25	0.302
1	70	90.38	20.38	415.34	4.595
2	10	6.78	3.22	10.37	1.529
3	2	0.34	3.65	13.32	37.733
4	1	0.013	—	—	—
5	1	0	—	—	—
				44.159	

The table value of $\chi^2 = 5.99$ at $\alpha = 0.05$ and $df = 6 - 1 - 3 = 2$. Since calculated value of $\chi^2 = 37.422$ is greater than its table value, the null hypothesis is rejected.

15.5.3 Yate's Correction for Continuity

The distribution of χ^2 -test statistic is continuous but the data under test is categorical which is discrete. It obviously causes errors, but it is not serious unless we have one degree of freedom, as in a 2×2 contingency table. To remove the probability of such errors to occur due to the effect of discrete data, we apply **Yate's correction** for continuity.

The correction factor suggested by Yate in case of a 2×2 contingency table is as follows:

- (a) Decrease by half those cell frequencies which are greater than expected frequencies and increase by half those which are less than expected.

This correction does not affect the row and column totals. For example, in a 2×2 contingency table where frequencies are arranged as:

Attributes	A	Not A	Total
B	a	b	$a + b$
Not B	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d = N$

The value of χ^2 calculated from independent frequencies is given by

$$\chi^2 = \frac{(a + b + c + d)(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$$

This value of χ^2 is corrected as:

$$\chi^2_{\text{corrected}} = \frac{n \left(ad - bc - \frac{1}{2}n \right)^2}{(a + b)(c + d)(a + c)(b + d)} ; \quad ad - bc > 0$$

$$\text{and} \quad \chi^2_{\text{corrected}} = \frac{n \left(bc - ad - \frac{1}{2}n \right)^2}{(a + b)(c + d)(a + c)(b + d)} ; \quad ad - bc < 0$$

Yate's correction: A continuity correction made when calculating the χ^2 -test statistic for a 2×2 contingency table.

(b) An alternative formula for calculating the χ^2 test statistic is as follows:

$$\chi^2 = \sum \frac{\{O - E\}^2}{E}$$

Example 15.8: Of the 1000 workers in a factory exposed to an epidemic, 700 in all were attacked, 400 had been inoculated, and of these 200 were attacked. On the basis of this information, can it be believed that inoculation and attack are independent?

[HP Univ., MBA, 1998]

Solution: Let us assume the null hypothesis that there is no association between inoculation and attack that is, inoculation and attack are independent.

The given information can be arranged in a 2×2 contingency table as follows:

Attributes	Attacked	Not Attacked	Total
Innoculated	200	200	400
Not inoculated	500	100	600
Total	700	300	1000

Using the result of Section 15.4.3, we have $a = 200$, $b = 200$, $c = 500$, $d = 100$, and $n = 1000$. Thus

$$\begin{aligned}\chi^2 &= \frac{(a+b+c+d)(ab-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \\ &= \frac{1000(200 \times 100 - 200 \times 500)^2}{(200+200)(500+100)(200+500)(200+100)} \\ &= \frac{1000 \times (80,000)^2}{400 \times 600 \times 700 \times 300} = \frac{1000 \times 64}{504} = 126.984\end{aligned}$$

The critical value of $\chi^2 = 3.84$ at $\alpha = 0.05$ and $df = (2-1)(2-1) = 1$. Since calculated value $\chi^2_{\text{cal}} = 126.984$ is more than its critical value, the null hypothesis rejected. Hence, we conclude that inoculation and attack are not independent.

Example 15.9: The following information was obtained in a sample of 40 small general shops:

Owner	Shops in		Total
	Urban Areas	Rural Areas	
Men	17	18	35
Women	2	12	15
Total	20	30	50

Can it be said that there are relatively more women owners of small general shops in rural than in urban areas?

Solution: Let us assume the null hypothesis that there are an equal number of men and women owners of small shops in both rural and urban areas.

Since one of the frequencies is small, we apply Yate's correction formula to calculate χ^2 as follows: Here $a = 17$, $b = 18$, $c = 3$, $d = 12$, and $n = 50$

$$\begin{aligned}\chi^2_{\text{corrected}} &= \frac{n \left(ad - bc - \frac{1}{2}n \right)^2}{(a+b)(c+d)(a+c)(b+d)} \\ &= \frac{50 \left(17 \times 12 - 18 \times 3 - \frac{1}{2} \times 50 \right)^2}{35 \times 15 \times 20 \times 30} \\ &= \frac{50 (204 - 54 - 25)^2}{35 \times 15 \times 20 \times 30} = \frac{50 \times 125 \times 125}{35 \times 15 \times 20 \times 30} = 2.480\end{aligned}$$

The critical value of $\chi^2 = 3.841$ at $\alpha = 0.05$ and $df = (2 - 1)(2 - 1) = 1$. Since calculated value $\chi^2 = 2.480$ is less than its critical value $\chi^2 = 3.841$, the null hypothesis is accepted. Hence we conclude that shops owned by men and women in both areas are equal in number.

Self-Practice Problems 15B

- 15.12** A sample analysis of examination results of 500 students was made. It was found that 220 students had failed, 170 had secured a third class, 90 were placed in second class, and 20 got a first class. Are these figures commensurate with the general examination result which is in the ratio of 4 : 3 : 2 : 1 for the various categories respectively. [Delhi Univ., MCom, 1999]

- 15.13** A set of 5 coins is tossed 3200 times, and the number of heads appearing each time is noted. The results are given below:

No. of heads :	0	1	2	3	4	5
Frequency :	80	570	1100	900	50	50

Test the hypothesis that the coins are unbiased.

[Annamalai Univ., MCom, 1998; MD Univ., MCom, 1998]

- 15.14** The demand for a particular spare part in a factory was found to vary from day to day. In a sample study the following information was obtained:

Day	:	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
No. of parts demanded	:	1124	1125	1110	1120	1126	1115

Test the hypothesis that the number of parts demanded does not depend on the day of the week.

[Delhi Univ., MBA, 2002]

- 15.15** The number of scooter accidents per week in a certain town were as follows:

12 8 20 2 14 10 15 6 9 4

Are these frequencies in agreement with the belief that accident conditions were the same during this 10-week period?

[Delhi Univ., MBA, 2000; HP Univ., MA Econ, 1997]

- 15.16** The divisional manager of a chain of retail stores believes the average number of customers entering each of the five stores in his division weekly is the same. In a given week, a manager reports the following number of customers in his stores : 3000, 2960, 3100, 2780, 3160. Test the divisional manager's belief at the 10 per cent level of significance [Delhi Univ., MBA, 2001]

- 15.17** Figures given below are (a) the theoretical frequencies of a distribution and (b) the frequencies of the Poisson distribution having the same mean and total frequency as in (a).

(a)	305	365	210	80	28	9	3
(b)	301	361	217	88	26	6	1

Apply the χ^2 test for goodness-of-fit.

Hints and Answers

- 15.12** Let H_0 : No difference in observed and expected results.

Category	O	E	$(O - E)^2$	$(O - E)^2/E$
Failed	220	$500(4/10) = 200$	400	2.000
3rd class	170	$500(3/10) = 150$	400	2.667
2nd class	90	$500(2/10) = 100$	100	1.000
1st class	20	$500(1/10) = 50$	900	18.000
			23.667	

Since $\chi^2_{\text{cal}} = 23.667$ is more than its critical value $\chi^2 = 7.81$ for $df = 4 - 1 = 3$ and $\alpha = 0.05$, the null hypothesis is rejected.

- 15.13** Let H_0 : Coins are unbiased that is, $p = q = 1/2$.

Apply binomial probability distribution to get the expected number of heads as follows:

Expected number of heads = $n^n C_r p^r q^{n-r}$

$$= 3200 \cdot {}^5 C_0 \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{5-r} = 3200 \cdot {}^5 C_0 \left(\frac{1}{2}\right)^5$$

O	E	$(O - E)^2$	$(O - E)^2/E$
80	100	400	4.00
570	500	4900	9.80
1100	1000	10,000	10.00
900	1000	10,000	10.00
500	500	0	0.00
50	100	2500	25.00
			58.80

Since $\chi^2_{\text{cal}} = 58.80$ is more than its critical value $\chi^2 = 11.07$ for $df = 6 - 1 = 5$ and $\alpha = 0.05$, the null hypothesis is rejected.

- 15.14** Let H_0 : Number of parts demanded does not depend on the day of the week.

Expected number of parts demanded = $6720/6 = 1120$ when all days are considered same.

Days	O	E	$(O - E)^2$	$(O - E)^2/E$
Monday	1124	1120	16	0.014
Tuesday	1125	1120	25	0.022
Wednesday	1110	1120	100	0.089
Thursday	1120	1120	0	0.000
Friday	1126	1120	36	0.032
Saturday	1115	1120	25	0.022
				0.179

Since $\chi_{\text{cal}}^2 = 0.179$ is less than its critical value $\chi^2 = 11.07$ for $df = 6 - 1 = 5$ and $\alpha = 0.05$, the null hypothesis is accepted.

- 15.15** Let H_0 : Accident conditions were same during the period.

Expected number of accidents per week = $(10 + 8 + 20 + \dots + 4)/10 = 10$

O	E	$(O - E)^2$	$(O - E)^2/E$
12	10	4	0.40
8	10	4	0.40
20	10	100	10.00
2	10	64	6.40
14	10	16	1.60
10	10	0	0.00
15	10	25	2.50
6	10	16	1.60
9	10	1	0.10
4	10	36	3.60
			26.60

Since $\chi_{\text{cal}}^2 = 26.60$ is more than its critical value $\chi^2 = 16.819$ for $df = 10 - 1 = 9$ and $\alpha = 0.05$, the null hypothesis is rejected.

- 15.16** Let H_0 : No significant difference in the number of customers entering each of the five stores.

Expected frequency of customers entering each store is $15,000/5 = 3000$.

O	E	$(O - E)^2$	$(O - E)^2/E$
3000	3000	0	0
2960	3000	1600	0.533
3100	3000	10,000	3.333
2780	3000	48,400	16.133
3160	3000	25,600	8.533
			28.532

Since $\chi_{\text{cal}}^2 = 28.532$ is more than its critical value $\chi^2 = 13.277$ for $df = 5 - 1 = 4$ at $\alpha = 0.05$, the null hypothesis is rejected.

- 15.17** Let H_0 : The Poisson distribution is a good fit to the given data

O :	305	365	210	80	28	<u>9</u>	3
E :	301	361	217	88	26	<u>6</u>	1

$$\chi^2 = \sum (O - E)^2/E = 4.8; df = 7 - 1 - 2 = 4; \alpha = 0.05;$$

$\chi_{\text{cal}}^2 = 9.49 > \chi_{\text{cal}}^2$, the H_0 is accepted.

15.5.4 χ^2 Test for Population Variance

The assumption underlying the χ^2 -test is that the population from which the samples are drawn is normally distributed. Let the variance of normal population be σ^2 . The null hypothesis is setup as: $H_0 : \sigma^2 = \sigma_0^2$, where σ_0^2 is hypothesized value of σ^2 .

If a sample of size n is drawn from this normal population, then variance of sampling distribution of mean \bar{x} is given by $s^2 = \sum (x - \bar{x})^2/(n - 1)$. Consequently the value of χ^2 -test statistic is determined as

$$\chi^2 = \frac{1}{\sigma^2} \sum (x - \bar{x})^2 = \frac{(n-1)s^2}{\sigma^2}$$

with $df = n - 1$ degrees of freedom.

Decision rule • If $\chi_{\text{cal}}^2 < \chi_{\alpha/2}^2$, then accept H_0

• Otherwise reject H_0

Confidence Interval for Variance We can define 95 per cent, 99 per cent, and so on. Confidence limits and intervals for χ^2 test statistic using table values of χ^2 . Same as t -distribution. Such limits of confidence helps to estimate population standard deviation in forms of sample standard deviation s with $(1 - \alpha)$ per cent confidence as follows:

$$\text{or } \frac{(n-1)}{\chi_{df, U}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{df, L}^2}$$

$$\text{or } \sqrt{\frac{(n-1)s^2}{\chi_{df, U}^2}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{\chi_{df, L}^2}}$$

$$\text{or } \frac{s\sqrt{n-1}}{\chi_{df, U}^2} \leq \sigma \leq \frac{s\sqrt{n-1}}{\chi_{df, L}^2}$$

where subscripts U and L stands for upper and lower tails proportion of area under χ^2 -curve. For example for a 95 per cent confidence interval the lowest 2.5 per cent and highest 25 per cent of χ^2 distribution curve is excluded leaving the middle 95 per cent area.

The χ^2 for upper 2.5 per cent ($= 0.025$) area is obtained directly for standard χ^2 table. To obtain value of χ^2 for lower 2.5 per cent ($= 0.025$) area, look under the 0.975 column of χ^2 table for given df , because $1 - 0.975 = 0.025$.

In case χ^2 table values are not available for larger df (≥ 30) values, then the values of $\chi_{df, v}^2$ for an appropriate confidence interval based on normal approximation can be obtained follows:

$$\chi_{df, U}^2 = \mu + z\sigma(\chi^2) \quad \text{and} \quad \chi_{df, L}^2 = \mu - z\sigma(\chi^2)$$

where, mean $\mu(\chi^2) = df$ and variance $\sigma^2(\chi^2) = 2df$

$$\text{and } \sigma(\chi^2) = \sqrt{v(\chi^2)}$$

Example 15.10: A random sample of size 25 from a population gives the sample standard deviation of 8.5. Test the hypothesis that the population standard deviation is 10.

Solution: Let us assume the null hypothesis that the population standard deviation $\sigma = 10$.

Given that $n = 25$; $s = 8.5$. Applying the χ^2 test statistic, we have

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(25-1)(8.5)^2}{(10)^2} = \frac{24(72.25)}{100} = 17.34$$

The critical value of $\chi^2 = 36.415$ at $df = 25 - 1 = 24$ and $\alpha = 0.05$. Since calculated of χ^2 is less than its critical value, the null hypothesis is accepted. Hence we conclude that the population variance is 10.

Example 15.11: A as a sample of 8 units from a normal population gives an unbiased estimate of population variance as 4.4. Find the 95 per cent confidence limits for population standard deviation σ .

Solution: Given that $\alpha = 5$ per cent, $n = 8$, $s^2 = 4.4$, $df = 8 - 1 = 7$. The confidence limits for σ^2 are as follows:

$$\begin{aligned} \frac{(n-1)s^2}{\chi_{df, U}^2} &< \sigma^2 < \frac{(n-1)s^2}{\chi_{df, L}^2} \\ \frac{(8-1)4.4}{16.01} &< \sigma^2 < \frac{(8-1)4.4}{1.69} \text{ or } 1.923 \leq \sigma^2 \leq 18.224 \end{aligned}$$

Consequently $1.386 \leq \sigma \leq 4.269$

Example 15.12: The standard deviation of lifetime of a sample of electric light bulbs is 100 hours. Find the 95 per cent confidence limits for the population standard deviation for such electric bulbs.

Solution: Given $\alpha = 5$ per cent, $n = 200$, $s = 100$, $df = n - 1 = 199$. The approximate 95 per cent confidence interval based on normal approximation requires that χ^2 values be approximated, where

$$\mu(\chi^2) = df = 199, \sigma^2(\chi^2) = 2df = 398, \text{ and } \sigma(\chi^2) = 19.949$$

Thus the approximate χ^2 values are

$$\chi_{df, U=0.025}^2 = \mu(\chi^2) + z\sigma(\chi^2) = 199 + 1.96(19.949) = 238.10$$

$$\chi_{df, L=0.025}^2 = \mu(\chi^2) - z\sigma(\chi^2) = 199 - 1.96(19.949) = 159.90$$

The 95 per cent confidence interval for the population standard deviation based on normal approximation of χ^2 values is given by

$$\sqrt{\frac{(n-1)s^2}{\chi_{df, U}^2}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{\chi_{df, L}^2}}$$

$$\frac{100\sqrt{199}}{15.430} \leq \sigma \leq \frac{100\sqrt{199}}{12.645} = 91.42 \leq \sigma \leq 111.55$$

Hence we can be 95 per cent confident that the population standard deviation will be between 91.42 hours and 111.55 hours.

15.5.5 Coefficient of Contingency

If a null hypothesis of independence is rejected at a certain level of significance, then it implies dependence of attributes on each other. In such a case we need to determine the measure of dependence in terms of association or relationship. The measure of the degree of relationship (association) of attributes in a contingency table is given by

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{s - n}{s}} ; \quad s = \sum (O/E)^2$$

This value of C is called the *coefficient of contingency*. A large value of C represents a greater degree of dependence or association between two attributes.

The maximum value which the coefficient of contingency C can take is $\sqrt{(k-1)/k}$, where k represents the number of rows and columns in the contingency table, such that $C = r = k$. When there is perfect dependence and $C = r = k$, the non-zero observed cell frequencies will occur diagonally and the calculated value of χ^2 will be as large as the sample size.

Example 15.13: 1000 students at college level were graded according to their IQ level and the economic condition of their parents.

	Economic Condition		IQ Level		Total
	High	Low			
Rich	460	140		600	
Poor	240	160		400	
Total	700	300		1000	

Use the coefficient of contingency to determine the amount of association between economic condition and IQ level.

Solution: Calculations for expected frequencies are as shown below:

O	E	O - E	$(O - E)^2$	$(O - E)^2/E$
460	420	40	1600	3.810
240	280	-40	1600	5.714
140	180	-40	1600	8.889
160	120	40	1600	<u>13.333</u>
				31.746

The coefficient of contingency, $C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{31.746}{31.746 + 1000}} = 0.175$ implies the level of association between two attributes, IQ level and economic condition.

The maximum value of C for a 2×2 contingency table is

$$C_{\max} = \sqrt{\frac{k-1}{k}} = \sqrt{\frac{2-1}{2}} = \sqrt{\frac{1}{2}} = 0.7071$$

which implies perfect dependence between the IQ level and economic condition of students.

15.5.6 Chi-Square Test of Homogeneity

The test of homogeneity is useful in a case when we intend to verify whether several populations are homogeneous with respect to some characteristic of interest. For example, we may like to know that the milk supplied by various companies has a particular ingredient

in common or not. Hence, the test of homogeneity is useful in testing a null hypothesis that several populations are homogeneous with respect to a characteristic.

This test is different from the test of independence on account of the following reasons:

- (i) Instead of knowing whether two attributes are independent or not, we may like to know whether different samples come from the same population.
- (ii) Instead of taking only one sample, for this test two or more independent samples are drawn from each population.
- (iii) When the characteristics to be compared consist of two categories, this test is similar to the test of hypothesis of difference between two population means or proportions.

To apply this test, first a random sample is drawn from each population, and then in each sample the proportion falling into each category or strata is determined. The sample data so obtained is arranged in a contingency table. The procedure for testing of hypothesis is same as discussed for test of independence.

Example 15.14: A movie producer is bringing out a new movie. In order to develop an advertising strategy, he wants to determine whether the movie will appeal most to a particular age group or whether it will appeal equally to all age groups. The producer takes a random sample from persons attending the preview of the new movie, and obtains the following results:

Opinion	Age Groups				Total
	Under 20	20-39	40-59	60 and over	
Liked the movie	146	78	48	28	300
Disliked the movie	54	22	42	22	140
Indifferent	20	10	10	20	60
Total	220	110	100	70	500

What inference will you draw from this data?

Solution: Let us assume the null hypothesis that the opinion of all age groups is same about the new movie.

The calculations for expected frequencies are as follow and displayed in the table below:

$$E_{11} = \frac{300 \times 220}{500} = 132, \quad E_{12} = \frac{300 \times 110}{500} = 66 \quad E_{13} = \frac{300 \times 100}{500} = 60$$

$$E_{14} = 300 - (132 + 66 + 60)$$

$$E_{21} = \frac{140 \times 220}{500} = 61.6, \quad E_{22} = \frac{140 \times 110}{500} = 30.8 \quad E_{23} = \frac{140 \times 100}{500} = 28$$

$$E_{24} = 140 - (61.6 + 30.8 + 28) = 42$$

$$E_{31} = 220 - (132 + 61.6) = 26.4 \quad E_{32} = 110 - (66 + 30.8) = 13.2$$

$$E_{33} = 100 - (60 + 28) = 12 \quad E_{34} = 70 - (42 + 19.6) = 8.4 = 19.6$$

O	E	O-E	(O-E) ²	(O-E) ² /E
146	132.0	14.0	196.00	1.485
54	61.6	-7.6	57.76	0.938
20	26.4	-6.4	40.96	1.552
78	66.0	12.0	144.00	2.182
22	30.8	-8.8	77.44	2.514
10	13.2	-3.2	10.24	0.776
48	60.0	-12.0	144.00	2.400
42	28.0	14.0	196.00	7.000
10	12.0	-2.0	4.00	0.333
28	42.0	-14.0	196.00	4.667
22	19.6	2.4	5.76	0.294
20	8.4	11.6	134.56	16.019
				40.16

The critical value of $\chi^2 = 12.59$ for $df = (r - 1)(c - 1) = (3 - 1)(4 - 1) = 6$ and $\alpha = 0.05$. Since the calculated value of χ^2 is greater than its critical, the null hypothesis is rejected. Hence we conclude that the opinion of all age groups is not the same about the new movie.

Conceptual Questions 15A

1. Describe the nature of a situation that gives rise to the multinomial distribution.
2. Consider a multinomial experiment in which the outcomes are classified into n categories. Explain why there are $n - 1$ degrees of freedom when using the chi-square goodness-of-fit procedure.
3. In the application of the chi-square procedure, describe how the expected frequencies are determined.
4. What is the χ^2 -test? Under what conditions is it applicable? Point out its role in business decision-making.
[Kumaon Univ., MBA, 2000]
5. Describe the χ^2 -test of significance and state the various uses to which it can be put.
6. What is the χ^2 -test of goodness-of-fit? What cautions are necessary while applying this test?
[Sukhadia Univ., MBA, 1998]
7. Explain Yate's method of correction for small frequencies in a contingency table.
8. Explain how the χ^2 -test is used in the test of homogeneity.
9. Under what conditions should the χ^2 -test of independence be used?
10. What are the advantages and limitations of the chi-square test of association? What is the general rule governing the applicability of the chi-square test?
11. What are similarities and differences between the z -test and the chi-square test?
12. Write a short note on each of the following:
 - (a) Chi-square statistic
 - (b) Critical value of chi-square
 - (c) Chi-square distribution
 - (d) Degrees of freedom

15.6 THE SIGN TEST FOR PAIRED DATA

This test is also known as **paired-sample sign test** and based upon the sign of difference in paired observations, say (x, y) where x is the value of an observation from population 1 and y is the value of an observation from population 2. This test assumes that the pairs (x, y) of values are independent and that the measurement scale within each pair is at least ordinal.

For comparing two populations, the sign test is stated in terms of probability that values from population 1 are greater than values from population 2, that are paired. The direction of the difference in values whether plus (+) or minus (-) is also recorded for each pair.

The probability (p) that a value x from population 1 will be greater than y , a value from population 2, is denoted by $p = P(x > y)$. Every pair of values in a sample is written as (x, y) , where $x > y$ is denoted by a plus (+) sign and $x < y$ is denoted by a minus (-) sign. Possibility that $x = y$ is ignored and is denoted by 0 (zero).

Since in the sign test ordering rather than actual measurements are involved, this test is acceptable when distribution is not symmetrical, i.e. when the mean would not be an appropriate measure of centerality.

Null and Alternative Hypotheses

In sign test, the null hypothesis, H_0 is stated that the probability of a plus (+) sign is equal to the probability of a minus (-) sign and both are 0.50, i.e. there is no difference between two populations. This situation is very similar to the fair-coin toss, where we use binomial distribution to describe sampling distribution. The possible hypotheses for the sign test are stated as follows:

One-tailed Test	Two-tailed Test
<ul style="list-style-type: none"> Right-tailed Test $H_0 : p \leq 0.50$ $H_1 : p > 0.5$ 	<ul style="list-style-type: none"> $H_0 : p = 0.50 \leftarrow$ No difference between two types of events $H_1 : p \neq 0.50 \leftarrow$ There is a difference between two types of events.
<ul style="list-style-type: none"> Left-tailed Test $H_0 : p \geq 0.5$ $H_1 : p < 0.5$ 	

The test-statistic to test the null hypothesis is defined as:

$$T = \text{Number of plus signs}$$

Since only two signs are considered and the probability (p) of getting a plus (+) sign is same as probability ($1 - p$) of getting a minus (-) sign, i.e. 0.50, therefore binomial distribution properties can be used to calculate expected number of plus (+) signs and possible variance. The binomial probability distribution can be approximated to a normal distribution if the conditions:

$$np \geq 5 \quad \text{and} \quad n(1-p) \geq 5$$

are satisfied. We can apply z-test statistic to test the null hypothesis $H_0 : p = 0.50$, so that

$$z = \frac{x - \mu}{\sigma} = \frac{x - np}{\sqrt{npq}}$$

where x is the number of possible signs observed.

Decision rule: If calculated value, z_{cal} of z-test statistic is less than its critical value, accept the null hypothesis. Otherwise reject H_0 .

Remark. The sign test can also be used to test the hypothesis that the median difference between two populations is zero. The null and alternative hypotheses are stated as:

$$H_0 : \text{Population median} = A, \text{ and } H_1 : \text{Population median} \neq A$$

where A is some number.

To conduct the sign test, observations are paired with the null hypothesis value of the median, and rest of the procedure remains same as discussed before.

Example 15.15: The nutritionists and medical doctors have always believed that vitamin C is highly effective in reducing the incidents of cold. To test this belief, a random sample of 13 persons is selected and they are given large daily doses of vitamin C under medical supervision over a period of one year. The number of persons who catch cold during the year is recorded and a comparison is made with the number of cold contacted by each such person during the previous year. This comparison is recorded as follows, along with the sign of the change.

Observations	:	1	2	3	4	5	6	7	8	9	10	11	12	13
Without vitamin C	:	7	5	2	3	8	2	4	4	3	7	6	2	10
With vitamin C	:	2	1	0	1	3	2	3	5	1	4	4	3	4
Sign	:	-	-	-	-	-	0	-	+	-	-	-	+	-

Using the sign test at $\alpha = 0.05$ level of significance, test whether vitamin C is effective in reducing colds.

Solution: Let us take the null hypothesis that there is no difference in the number of cold contacted with or without vitamin C and this probability (p) = 0.50, i.e.

$$H_0 : p = 0.50 \quad \text{and} \quad H_1 : p \neq 0.50$$

Given $n = 12$ (because difference is zero in observation 6); number of minus signs = 10 and number of plus signs = 2. Thus

$$\mu = np = 12(0.5) = 6 \text{ and } \sigma = \sqrt{npq} = \sqrt{12 \times 0.5 \times 0.5} = \sqrt{3} = 1.73$$

Applying the z-test statistic; we get

$$z = \frac{\bar{x} - \mu}{\sigma} = \frac{9.5 - 6}{1.73} = 2.02$$

where, $\bar{x} = 10 \approx 9.5$ because of approximation from discrete distribution to normal distribution.

Since z_{cal} ($= 2.02$) is greater than the critical value z_α ($= 1.96$) at $\alpha = 0.05$ level of significance, H_0 is rejected. Hence we conclude that there is a significant difference in the number of cold contacted with or without vitamin C.

Example 15.16: The median age of tourists who has come to India is claimed to be 40 years. A random sample of 18 tourists gives the following ages:

24, 18, 37, 51, 56, 38, 45, 29, 48, 39, 26, 38, 43, 62, 30, 66, 41

Test the hypothesis using $\alpha = 0.05$ level of significance.

Solution: Let us take the null and alternative hypotheses as stated below:

$$H_0: \mu = 40 \quad \text{and} \quad H_1: \mu \neq 40$$

Arranging data on ages of tourists in an ascending order and making pair of those with median age 40 years to determine plus (+) sign and minus (-) sign as follows:

Tourist Age :	18	24	26	29	30	37	38	38	39	41	43	45	45	48	51	56	62	66
Median age :	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40
Sign	:	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-

Given $n=18$, number of plus (+) sign, $x=9$, and $p=0.5$. Thus

$$\mu=np=18(0.5)=9, \text{ and } \sigma=\sqrt{npq}=\sqrt{18\times 0.5\times 0.5}=1.060$$

Applying the z -test statistics, we get

$$z = \frac{\bar{x} - \mu}{\sigma} = \frac{8.5 - 9}{1.060} = -1.533$$

where, $\bar{x} = 9 \cong 8.5$ because of approximation from discrete distribution to normal distribution.

Since z_{cal} ($= -1.533$) is greater than its critical value z_α ($= -1.96$) at $\alpha = 0.05$ level of significance, H_0 is accepted. Hence, we conclude that the claim is correct.

15.7 RUNS TEST FOR RANDOMNESS

The randomness of the sample drawn from a population is essential for all types of statistical testing because sample results are to be used to draw conclusions regarding the population under study. The *run test* helps to determine whether the order or sequence of observations (symbols, items or number) in a sample is random. The runs test examines the number of 'runs' of each of two possible characteristics that sample elements may have. *A run is a sequence of identical occurrences of elements (symbols or numbers) preceded and followed by different occurrences of elements or by no element at all.* For example, in tossing a coin, the outcome of three tails in succession would constitute a run, as would a succession of five heads. To quantify how many runs are acceptable before raising doubt about the randomness of the process, a probability distribution is used that leads to a *statistical test for randomness*.

Suppose that, tossing a coin 20 times produces the following sequence of heads (H) and tails (T).

HHH	TT	HH	TTT	HHH	TT	HHH	T
1	2	3	4	5	6	7	8

In this example, the first run of HHH is considered as run 1, the second run of TT as run 2 and so on, so that there are 8 runs in all or in other words, $r = 8$. However, in this example, rather than perfect separation between H and T, it appears to be a perfect clustering together. It is a form of regularity not likely to have arisen by chance.

Small Sample Run Test

In order to test the randomness, let n_1 = number of elements of one kind, and n_2 = number of elements of second kind. Total sample size is $n = n_1 + n_2$. In the above example, $n_1 = 12$ heads and $n_2 = 8$ tails. Let one kind of elements be denoted by plus (+) sign and second kind of elements be denoted by minus (-) sign. In the above example, H is represented by plus (+) sign and T by minus (-) sign. The concept of plus (+) or minus (-) provides the direction of change from an established pattern. Accordingly, a plus (+) would be considered a change from an established pattern value in one direction and a minus (-) would be considered a change in the other direction.

If the sample size is small, so that either n_1 or n_2 is less than 20, then test is carried out by comparing the deserved number of runs R to critical values of runs for the given values of n_1 and n_2 . The critical values of R are given in Appendix. The null and alternative hypotheses stated as:

H_0 : Observations in the samples are randomly generated

H_1 : Observations in the samples are not randomly generated

can be tested that the occurrences of plus (+) signs and minus (-) signs are random by comparing r value with its critical value at a particular level of significance. **Decision rule**

- Decision rule:**
- Reject H_0 at a if $R \leq C_1$ or $R \geq C_2$
 - Otherwise accept H_0

where C_1 and C_2 are critical values obtained from standard table with total tail probability $P(R \leq C_1) + P(R \geq C_2) = \alpha$.

Large Sample Run Test

If the sample size is large so that either n_1 or n_2 is more than 20, then the sampling distribution of R statistic (i.e. run) can be closely approximated by the normal distribution. The mean and standard deviation of the number of runs for the normal distribution are given by

$$\text{Mean, } \mu_R = \frac{2n_1 n_2}{n_1 + n_2} + 1$$

$$\text{Standard deviation, } \sigma_R = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}$$

Thus the standard normal test statistic is given by

$$z = \frac{R - \mu_R}{\sigma_R} = \frac{r - \mu_R}{\sigma_R}$$

The critical z-value is obtained in the usual manner at a specified level of significance.

Example 15.17: A stock broker is interested to know whether the daily movement of a particular share averages in the stock market showed a pattern of movement or whether these movements were purely random. For 14 business days, he noted the value of this average and compared it with the value at the close of the previous day. He noted the increase as plus (+) and decrease as minus (-). The record was as follows:

+, +, -, -, +, +, +, -, +, +, -, +, -, -

Test whether the distribution of these movements is random or not at $\alpha = 0.05$ level of significance.

Solution: Let us state the null and alternative hypotheses as follows:

H_0 : Movement is random and H_1 : Movement is not random

There are $r = 8$ runs (plus signs) with $n_1 = 8$ plus (number of increases) and $n_2 = 6$ minus (number of decreases) so that $n = n_1 + n_2 = 14$. The critical value of $r = 8$ for $n_1 = 8$ and $n_2 = 6$, implies that H_0 is rejected when $r \leq 3$ and $r \geq 12$ at $\alpha = 0.05$ level of significance. Since $3 \leq r \leq 12$, therefore H_0 cannot be rejected.

Example 15.18: Some items produced by a machine are defective. If the machine follows some pattern where defective items are not randomly produced throughout the process the machine needs to be adjusted. A quality control engineer wants to determine whether the sequence of defective (D) versus good (G) items is random. The data are

GGGGG, DDD, GGGGGG, DDD, GGGGGGGGGG, DDDD,
GGGGGGGGGG, DDD, GGGGGGGGGG, DDDD

Test whether the distribution of defective and good items is random or not at $\alpha = 0.05$ level of significance.

Solution: Let us state the null and alternative hypotheses as follows:

H_0 : Sequence is random, and H_1 : Sequence is not random

There are $r = 10$ runs with $n_1 = 43$ (number of G) and $n_2 = 17$ (number of D), so that $n = n_1 + n_2 = 60$. Since sample size is large, we compute mean and standard deviation as follows:

$$\text{Mean, } \mu_R = \frac{2n_1 n_2}{n_1 + n_2} + 1 = \frac{2 \times 43 \times 17}{43 + 17} + 1 = 24.8$$

$$\begin{aligned}\text{Standard deviation, } \sigma_R &= \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}} \\ &= \sqrt{\frac{2(43)(17)[2(43)(17) - 43 - 17]}{(43 + 17)^2 (43 + 17 - 1)}} \\ &= \sqrt{\frac{1462 \times 1402}{3600 \times 59}} = \sqrt{\frac{20,49,724}{2,12,400}} = 3.106\end{aligned}$$

The computed value of z-test statistic is

$$z = \frac{R - \mu_R}{\sigma_R} = \frac{10 - 24.8}{3.106} = -7.871$$

The critical value of $z_\alpha = 2.58$ for two-tailed test at $\alpha=0.05$ level of significance. Since $z_\alpha < z_{\text{cal}}$, H_0 is rejected. Hence, we conclude that the sequence of defective versus good item is not random.

15.8 MANN-WHITNEY U-TEST

If sample size is small and we cannot or do not wish to make the assumption that the data is taken from normally distributed population, then Mann-Whitney U-test or simply U-test is used to test the equality of two population means. This test is the substitute for t-test statistic when the stringent assumptions of parent population being normally distributed with equal variance are not met or when the data are only ordinal in measurement.

To apply the U-test, also called the **Wilcoxon rank sum test** or simply the **Rank Sum Test** values from two samples are combined into one group and are arranged in ascending order. The pooled values are then ranked from 1 to n with smallest value being assigned a rank 1. Such a test produces good results when data are on ordinal scale of measurement. The sum of ranks of values from sample 1 is denoted as R_1 . Similarly the sum of the ranks of values from sample 2 is denoted as R_2 . If both n_1 and $n_2 < 30$, samples are considered small. If either n_1 or n_2 is greater than 10, the samples are considered large.

Small Sample U-Test

- Combine the two random samples of size n_1 and n_2 and rank values from smallest to largest. If several values are tied, then assign each the average of the ranks that would otherwise have been assigned.

If the two sample sizes are unequal, then suppose n_1 represent smaller-sized sample and n_2 the larger-sized sample. The rank sum test statistic U_1 is the sum of the ranks assigned to the n_1 observations in the smaller sample. However, for equal-sized samples, either group may be selected for determining U_1 .

The test statistic U_1 plus the sum of the ranks assigned to the n_2 observations in the larger sample:

$$U_1 + U_2 = \frac{n(n+1)}{2}$$

represents the sum of first consecutive integers.

- State the null and alternative hypotheses for U-test as follows:

$H_0 : u_1 = u_2 \leftarrow$ Two populations distribution have equal mean

$H_1 : u_1 \neq u_2 \leftarrow$ Two populations distribution have different means

The test of the null hypothesis can either be two-tailed or one-tailed.

- The value of U-statistic is the smallest of the following two U-values computed as follows:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2$$

The U-statistic is a measure of the difference between the ranked observations of the two samples. Large or small values of statistic provide evidence of a difference between two populations. If differences between populations are only in location, then large or small values of U-statistic provide evidence of a difference in location (median) of two populations.

Both U_1 and U_2 need not be calculated, instead one of U_1 or U_2 can be calculated and other can be formed by using the equation: $U_1 = n_1 n_2 - U_2$.

Large Sample U-Test

For large samples (i.e. when both n_1 and n_2 are greater than 10) the sampling distribution of the U-statistic can be approximated by the normal distribution so that z-test statistic is given by

$$z = \frac{U - \mu_U}{\sigma_U}$$

where Mean $\mu_U = \frac{n_1 n_2}{2}$ and standard deviation, $\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$

Decision rules

- When n_1 and n_2 are both less than or equal to 10, standard table value can be used to obtain the critical value of the test statistic for both one and two tailed test at a specified level of significance.
- For large sample, at a specified level of significance.
 - Reject H_0 , if computed value of $z_{\text{cal}} \geq$ critical value z_α
 - Otherwise accept H_0

Example 15.19: It is generally believed that as people grow older, they find it harder to go to sleep. To test if there was a difference in time in minutes before people actually went to sleep after lying in the bed, a sample of 10 young persons (ages 21 to 25) and 10 old persons (ages 65 to 70) was randomly selected and their sleeping habits were monitored. The data show the number of minutes these 20 persons were awake in bed before getting to sleep:

Young men :	58	42	68	20	15	35	26	40	47	28
Old men :	100	152	147	70	40	95	68	90	112	58

Is there evidence that young men are significantly take more time to get to sleep than old men. Use $\alpha = 0.05$ level of significance.

Solution: First arrange the data in ascending order for ranking as shown below. When ties occur, we assign to each tied observation the average rank of the ties.

Young men :	15	20	26	28	35	40	42	47	58	68
Old men :	-	-	-	-	-	40	-	-	58	68
	70	90	95	100	112	147	152	-	-	-

List the ranks of all the observations in each of the two groups as follows:

Young men :	1	2	3	4	5	6.5	8	9	10.5	12.5	= 64.5
Old men :	6.5	10.5	12.5	14	15	16	17	18	19	20	= 148.5

Let us define young men as population 1 and old men as population 2. The rank sum are $R_1 = 64.5$ and $R_2 = 148.5$. Computing the test statistic U, we get

$$\begin{aligned} U_1 &= n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 = 10 \times 10 + \frac{10 \times 11}{2} - 64.5 \\ &= 100 + 55 - 64.5 = 90.5 \end{aligned}$$

$$\begin{aligned} U_2 &= n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2 = 10 \times 10 + \frac{10 \times 11}{2} - 148.5 \\ &= 100 + 55 - 148.5 = 6.5 \end{aligned}$$

Consider the lower value between U_1 and U_2 , i.e. $U_2 = 6.5$, so that

$$\text{Mean, } \mu_U = \frac{n_1 n_2}{2} = \frac{10 \times 10}{2} = 50$$

$$\text{and Standard deviation, } \sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{10 \times 10 (10 + 10 + 1)}{12}}$$

$$= \sqrt{\frac{100 \times 21}{12}} = \sqrt{\frac{2100}{12}} = 13.23.$$

Since both n_1 and n_2 are greater than 8, applying z -test statistic so that

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{6.5 - 50}{13.23} = -3.287$$

Since computed value of z_{cal} ($= -3.287$) is less than its critical value $z_\alpha = -1.96$ at $\alpha = 0.05$ significance level, null hypothesis is rejected. Hence we conclude that there is a significant difference in time to get to sleep between young men and old men.

15.9 WILCOXON MATCHED PAIRS TEST

This test is also known as **Wilcoxon signed rank test**. When two samples are related, the U-test discussed before is not applicable. Wilcoxon test is a non-parametric test alternative to t-test for two related samples. In U-test the differences between paired observations is not believed to be normally distributed and also ignore the size of magnitude of these differences. Wilcoxon test takes into consideration both the direction and magnitude of differences between paired values.

Procedure

- Compute differences in the same manner as in the case of the U-test. Then assign ranks for 1 to n to the absolute values of the differences starting from the smallest to largest differences. All pairs of values with zero differences are ignored. If differences are equal in magnitude, a rank equal to the average of ranks that would have been assigned otherwise is given to all the equal differences.
- Take the sum of the ranks of the positive and of the negative differences. The sum of positive and negative differences is denoted by s_+ and s_- respectively.
- Define Wilcoxon T-statistic as the smallest sum of ranks (either s_+ or s_-):

$$T = \min(s_+, s_-) = s$$

Since values of s_+ , s_- and s may vary in repeated sampling, these sums for a given sample may be treated as specific values of their respective sample statistic.

When the number of pairs of values, n is more than 15, the value of T is approximately normally distributed and z -test statistic is used to test the null hypothesis.

- Define null and alternative hypotheses as follows:

One-tailed Test	Two-tailed Test
$H_0 : \mu_1 = \mu_2 \text{ or } \mu_1 - \mu_2 = d_0$	$H_0 : \mu_1 = \mu_2 \text{ or } \mu_1 - \mu_2 = d_0$
$H_1 : \mu_1 > \mu_2 \text{ or } \mu_1 - \mu_2 > d_0$	$H_1 : \mu_1 \neq \mu_2 \text{ or } \mu_1 - \mu_2 \neq d_0$

- See table value of s_α for different sample size n and a specified level of significance α for s_+ , s_- and s_α . Decision rules for one-tailed and two-tailed test are as follows:

One-tailed Test	Two-tailed Test
<ul style="list-style-type: none"> • Reject H_0 when $s_- < s_\alpha$ or $s_+ < s_\alpha$ • Otherwise accept H_0 	<ul style="list-style-type: none"> • Reject H_0 when $s < s_\alpha$ • Otherwise accept H_0

Remarks

- If the sample size is small, i.e. $n \leq 15$ (number of pairs), then a critical value against which to compare T can be found by using n and α . If calculated value of T is less than or equal to the critical value of T at α level of significance and sample size n , then H_0 is rejected.

2. If the sample size n (> 15), then the sampling distribution of T (i.e. s_+ and s_-) approaches normal distribution with

$$\text{Mean } \mu_T = \frac{n(n+1)}{4}$$

$$\text{and Standard deviation, } \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

Given the level of significance α , z-test statistic is computed as:

$$z = \frac{T - \mu_T}{\sigma_T}$$

where n = number of pairs.

Example 15.20: Ten workers were given on-the-job training with a view to shorten their assembly time for a certain mechanism. The results of the time (in minutes) and motion studies before and after the training programme are given below:

Worker :	1	2	3	4	5	6	7	8	9	10
Before :	61	62	55	62	59	74	62	57	64	62
After :	59	63	52	54	59	70	67	65	59	71

Is there evidence that the training programme has shortened the average assembly time?

Solution: Let us take the null hypothesis that the training programme has helped in reducing the average assembly time. Calculations to compute Wilcoxon T-statistic are shown below:

Table 15.2: Computation of Wilcoxon T-statistic.

Worker	Before the Training (x_1)	After the Training (x_2)	Difference $d = x_1 - x_2$	Absolute Rank	Signed Rank	
					Positive (s_+)	Negative (s_-)
1	61	59	+ 2	2	2	-
2	62	63	- 1	1	-	1
3	55	52	+ 3	3	3	-
4	62	54	+ 8	7	7	-
5	59	59	0	Ignored	-	-
6	74	70	+ 4	4	4	-
7	62	67	- 5	5.5	-	5.5
8	57	45	+ 12	9	9	-
9	64	59	+ 5	5.5	5.5	-
10	62	71	- 9	8	-	8
				$\Sigma s_+ = 30.5$	$\Sigma s_- = 14.5$	

Since the smaller sum is associated with the negative ranks, value of the Wilcoxon test statistic is: $T = \Sigma s_- = 14.5$. We compare the computed value of T with its critical value $s = 8$ at $n = 10$ and $\alpha = 0.05$ level of significance (See Appendix). Since computed value of T is more than its critical value, null hypothesis is accepted and we conclude that the training has not helped in raducing the average assembly time.

Alternative approach

$$\mu_T = \frac{n(n+1)}{4} = \frac{10 \times 11}{4} = 27.5$$

$$\text{and } \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{9.81}{24}} = 9.81$$

Applying z-test statistic, we get

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{14.5 - 27.5}{9.81} = -1.325$$

Since computed value of z ($= -1.325$) is less than the critical value z_α ($= \pm 1.96$) at $\alpha = 0.05$ level of significance, null hypothesis is accepted.

Example 15.21: The average score on a vocational training test has been known to be 64. Recently, several changes have been carried out in the programme; the effect of these changes on performance on the test is unknown. It is therefore desirable to test the null hypothesis that the average score for all people who will complete the programme will be 64 versus the alternative that it will not be 64. The following random sample of scores is available

87, 91, 65, 31, 8, 53, 99, 44, 42, 60, 77, 73, 42, 50
79, 90, 54, 39, 77, 60, 33, 41, 42, 85, 71, 50

Is there evidence that the average score for all people who will complete the programme will be 64?

Solution: Let us take the null and alternative hypotheses as:

$$H_0 : \mu = 64 \quad \text{and} \quad H_1 : \mu \neq 64$$

Subtracting the average score of 64 (the null hypothesis mean) from every data point to form pairs. This gives the differences:

$$\begin{aligned} 87 - 64 &= +23, \quad 91 - 64 = +27, \quad 65 - 64 = +1, \quad 31 - 64 = -33, \quad 8 - 64 = -56, \\ 53 - 64 &= -11, \quad 99 - 64 = +35, \quad 44 - 64 = -20, \quad 42 - 64 = -22, \quad 60 - 64 = -4, \\ 77 - 60 &= +17, \quad 73 - 64 = -9, \quad 42 - 64 = -22, \quad 50 - 64 = -14, \quad 79 - 64 = +15, \\ 90 - 64 &= +26, \quad 54 - 64 = -10, \quad 39 - 64 = -25, \quad 77 - 64 = -13, \quad 60 - 64 = -4, \\ 33 - 64 &= -31, \quad 41 - 64 = -23, \quad 42 - 64 = -22, \quad 85 - 64 = +21, \quad 71 - 64 = +7, \\ 50 - 64 &= -14. \end{aligned}$$

Ranking the absolute value of the differences from smallest to largest, we get

Difference :	+23	+27	+1	-33	-56	-11	+35	-20	-22	-4	+17	+9	-22
Rank :	18.5	22	1	24	26	7	25	13	16	2.5	12	5	16
Difference :	+4	+15	+26	-10	-25	13	-4	-31	-23	-22	+21	+7	-14
Rank :	9.5	11	21	6	20	8	2.5	23	18.5	16	14	4	9.5
Σs_+ :	18.5 + 22 + 1 + 25 + 12 + 5 + 11 + 21 + 8 + 14 + 4 = 141.5												
Σs_- :	24 + 26 + 7 + 13 + 16 + 2.5 + 16 + 9.5 + 6 + 20 + 2.5 + 23 + 18.5 + 16 + 9.5 = 209.5												

Since the smaller sum is associated with positive ranks, we define T as that sum, i.e. $T = 141.5$.

Consider the sampling distribution of T-statistic as normal so that

$$\mu_T = \frac{n(n+1)}{4} = \frac{26 \times 27}{4} = 175.50$$

$$\text{and} \quad \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{26 \times 27 \times 53}{24}} = \sqrt{\frac{37206}{24}} = 39.37$$

Applying z-test statistic, we get

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{141.5 - 175.50}{39.37} = -0.863$$

Since computed value of z_{cal} ($= -0.863$) is less than the critical value $z_\alpha = \pm 1.96$ at $\alpha = 0.05$ level of significance, null hypothesis is accepted. Hence, we conclude that average score for all people who will complete the programme will be 64.

15.10 KRUSKAL-WALLIS TEST

This test is the non-parametric alternative to the one-way analysis of variance to identify differences among populations that does not require any assumption about the shape of the population distribution. This test uses the ranks of the observations rather than the data themselves with the assumption that the observations are on an interval scale.

Procedure: This test is used for comparing k different populations having identical distribution. The summary of procedure is as follows:

1. Draw k independent samples n_1, n_2, \dots, n_k from each of k different populations. Then combine these samples such that $n = n_1 + n_2 + \dots + n_k$, and arrange these n observations in an ascending order.
2. Assign ranks to observations from 1 to n such that smallest value is assigned rank 1. For ties each value is assigned average rank.
3. Identify rank values whether they belong to samples of size n_1, n_2, \dots, n_k . The ranks corresponding to different samples are totaled and the sum for the respective sample is denoted as t_1, t_2, \dots, t_k .

The following formula is used to compute Kruskal Wallis H-statistic:

$$H = \frac{12}{n(n+1)} \left[\sum_{j=1}^k \frac{t_j^2}{n_j} \right] - 3(n+1)$$

This H value is approximately Chi-square distributed with $k - 1$ degrees of freedom as long as n_j is not less than 5 observations for any population

However, if one or more samples have two or more equal observations, the value of H is adjusted as : $H' = H/C$, where C is the correction factor defined as:

$$C = 1 - \frac{1}{n^3 - n} \left[\sum_{j=1}^r (O_j^3 - O_j) \right]$$

where O_j = the number of equal observations in the j th sample.

r = the number of samples which has equal observations

The null and alternative hypotheses are stated as:

H_0 : The k different populations have identical distribution

H_1 : At least one of the k populations has different distribution.

Decision Rule

- Reject H_0 when computed value of H is greater than χ^2 (Chi-square) at $df = k - 1$ and α level of significance
- Otherwise accept H_0 .

Example 15.22: Use Kruskal-Wallis test to determine whether there is a significant difference in the following populations. Use $\alpha = 0.05$ level of significance

Population 1 : 17 19 27 20 35 40

Pupulation 2 : 28 36 33 22 27

Population 3 : 37 30 39 42 28 25 31

Solution: Three populations are considered for study, so $k = 3$ and $n = 18$. The observations in three populations are combined and ranked. The smallest value is given rank 1, as shown below:

Populatin 1		Populations 2		Population 3	
Value	Rank	Value	Rank	Value	Rank
17	1	22	4	25	5
19	2	27	6.5	28	8.5
20	3	28	8.5	30	10
27	6.5	33	12	31	11
35	13	36	14	37	15
40	17			39	16
$n_1 = 6$	$t_1 = 42.5$	$n_2 = 5$	$t_2 = 45$	$n_3 = 7$	$t_3 = 83.5$
				42	18

Suppose H_0 : Three populations are identical, i.e. $\mu_1 = \mu_2 = \mu_3$

$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$

Kruskal-Wallis test statistic is

$$H' = \frac{H}{C},$$

$$\begin{aligned}
 \text{where } H &= \frac{12}{n(n+1)} \left[\sum_{j=1}^3 \left(t_j^2 / n_j \right) \right] - 3(n+1) \\
 &= \frac{12}{18 \times 19} \left[\frac{(42.5)^2}{6} + \frac{(45)^2}{5} + \frac{(83.5)^2}{7} \right] - 3 \times 19 \\
 &= 0.035[301.4 + 405 + 996.03] - 57 = 2.572 \\
 \text{and } C &= 1 - \frac{1}{n^3 - n} \left[\sum_{j=1}^2 (O_j^3 - O_j) \right] \\
 &= 1 - \frac{1}{(18)^3 - 18} [\{(2)^3 - 2\} + \{(2)^2 - 2\}] \\
 &= 1 - \frac{12}{5814} = 1 - 0.002 = 0.998 \\
 \text{Thus } H' &= \frac{2.572}{0.998} = 2.577
 \end{aligned}$$

Since computed value of H' ($= 2.577$) is less than table value of χ^2 ($= 5.99$) at $df = k - 1 = 2$ and $\alpha = 0.05$, the null hypothesis is accepted and conclude that three populations are identical.

Self-Practice Problems 15C

- 15.18** Pre- and post-test scores after a particular training programme are known to be non-normal in their distribution. A sample of the scores, with the calculated changes, is given below:

Pre-test :	67	71	83	69	68	36	52	72	56
	64	76	83	69	68	36			
	52	72	56						
Post-test :	58	62	84	67	72	38	63	72	55
	59	76	84	69	72	38			
	63	74	66						

Conduct a sign test for determining whether any significant change has taken place.

- 15.19** The median age of a tourist to India is claimed to be 41 years. A random sample of 18 tourists gives the following ages:

25, 19, 38, 52, 57, 39, 46, 46, 30, 49, 40, 27, 39, 44, 63, 31, 67, 42

Test the hypothesis against a two-tailed alternative using $\alpha = 0.05$.

- 15.20** A courier service employs eight men and nine women. Every day, the assignments of delivery are supposed to be done at random. On a certain day, all the best jobs, in order of desirability, were given to the eight men. Is there evidence of sex discrimination? Discuss this also in the context of a continuing, daily operation. What would happen if you tested the randomness hypothesis everyday?

- 15.21** The owner of a garment shop wants to test whether his two salesmen A and B are equally effective. That is, he wants to test whether the number of sales made by each salesman is about the same or whether

one salesman is better than the other. He gets the following random samples of daily sales made by each salesman.

Salesman A : 35, 44, 39, 50, 48, 29, 60, 75, 49, 66

Salesman B : 17, 23, 13, 24, 33, 21, 18, 16, 32

Test whether salesperson A and B are equally effective.

- 15.22** Suppose 26 cola drinkers are sampled randomly to determine whether they prefer brand A or brand B. The random sample contains 18 drinkers of brand A and 8 drinkers of brand B. Let C denotes brand A drinkers and D denote brand B drinkers. Suppose the sequence of sampled cola drinkers is DCCCCCDCCDCDCCDCDC CCDDDCCC. Is this sequence of cola drinkers evidence that the sample is not random.

- 15.23** A machine produces parts that are occasionally flawed. When the machine is working in adjustment, flaws still occur but seem to happen randomly. A quality control person randomly selects 50 of the parts produced by the machine today and examines them one at a time in the order that they were made. The result is 40 parts with no flaws and 10 parts with flaws. The sequence of no flaws (denoted by N) and flaws (denoted by F) is shown here. Using $\alpha=0.05$ level of significance, the quality controller tests to determine whether the machine is producing randomly (the flaws are occurring randomly).

NNN F NNNNNNN F NN FF NNNNNN F
NNNN F NNNNNN FFFF NNNNNNNNNNNN

Is this sequence of flaws and no flaws evidence that the sample is not random?

- 15.24** A panel of 8 members has been asked about their perception of a product before and after they had an opportunity to try it. Their perceptions, measured on an ordinal scale, gave the results given below:

Member :	A	B	C	D	E	F	G	H
Before :	8	3	6	4	5	7	6	7
After :	9	4	4	1	6	7	9	2

Have the perception scores changed after trying the product? Use $\alpha = 0.05$ level of significance.

- 15.25** Samples have been taken from two branches of a chain of stores. The samples relate to the daily turnover of both the branches. Is there any difference in turnover between the two branches?

Branch 1 :	23500	25500	35500	19500
	24400	24000	23600	25900
	26000			
Branch 2 :	24000	19800	22000	21500
	24500			

- 15.26** The average hourly number of messages transmitted by a communications satellite is believed to be 149. If there is a possibility that demand for this service may be declining, then test the null hypothesis that the average hourly number of relayed messages is 149 (or more) versus the alternative hypothesis that the average hourly number of relayed messages is less than 149. A random sample of 25 operation hours is selected. The data (numbers of messengers relayed per hour) are 151, 144, 123, 178, 105, 112, 140, 167, 177, 185, 129, 160, 110, 170, 198, 165, 109, 118, 155, 102, 164, 180, 139, 166, 182.

Is there any evidence of declining the use of the satellite?

- 15.27** The average life of a 100-watt light bulb is stated on the package to be 750 hours. The quality control manager at the plant making the lightbulbs needs to check whether the statement is correct. The manager is only concerned about a possible reduction in quality and will stop the production process only if statistical evidence exists to conclude that the average life of a lightbulb is less than 750 hours. A random sample of 20 bulbs is collected and left on until they burn out. The lifetime of each bulb is recorded. The data are (in hours of continuous use) 738, 752, 710, 701, 689, 779, 650, 541, 902, 700, 488, 555, 870, 609, 745, 712, 881, 599, 659, 793. Should the process be stopped and corrected? Explain why or why not.

- 15.28** An accounting firm wants to find out whether the current ratio for three companies is same. Random samples of eight firms in industry A, six firms in industry B, and six firms in industry C are available. The current ratios are as follows:

Company A	Company B	Company C
1.38	2.33	1.06
1.55	2.50	1.37
1.90	2.79	1.09
2.00	3.01	1.65
1.22	1.99	1.44
2.11	2.45	1.11
1.98		
1.61		

Conduct the test at $\alpha = 0.05$ level of significance, and state your conclusion.

- 15.29** Results of a survey indicated that people between 55 and 65 years of age contact a physician an average of 9.8 times per year. People of age 66 and older contact doctors on an average of 12.9 times per year. Suppose you want to validate these results by taking your own samples. The following data represent the number of annual contacts people make with a physician. The samples are independent. Apply a suitable test statistic to determine whether the number of contacts with physicians by the people of age 65 years and older is greater than the number by people of age 55 to 65 years

55 to 65 : 12 13 8 11 9 6 11

65 and older : 16 15 10 17 13 12 14 9 13

- 15.30** Consider the survey that estimated the average annual household spending on healthcare. The metropolitan average was Rs.1800. Suppose six families in Delhi are matched demographically with six families in Mumbai and their amounts of household spending on healthcare for last year are obtained. The data are as follow:

Family Pair	Delhi	Mumbai
1	1950	1760
2	1840	1870
3	2015	1810
4	1580	1660
5	1790	1340
6	1925	1765

Apply a suitable test statistic to determine whether there is a significant difference in annual household healthcare spending between these two cities.

Hints and Answers

15.18 H_0 : No difference between the scores

H_1 : There is a difference between scores

$$s_+ = 10, s_- = 5, s = 5; \text{Accept } H_0$$

15.19 $T = 9, H_0$ is accepted.

15.20 $z = -3.756, H_0$ is rejected

15.21 $n_1 = 10$ and $n_2 = 9$. Arrange values in two samples in increasing order and denote them by A or B based on which population they come from:

BBBBBBBBBABBAAAAAA

Total runs $R = 4$. Salesman A sells more than B.

15.22 H_0 : The observations in the sample are random.

H_1 : The observations in the sample are not random.

Tally the number of runs

D	CCCCC	D	CC	D	CCCC
1	2	3	4	5	6
D	C	D	CCC	DDD	CCC
7	8	9	10	11	12

The number of runs, $R = 12$. Since the value of R falls between the critical values of 7 and 17, do not reject H_0 .

15.23 H_0 : The observations in the sample are random.

H_1 : The observations in the sample are not random.

Apply z-test statistic is

$$\mu_R = R - \left(\frac{2n_1 n_2}{n_1 + n_2} + 1 \right) = \frac{2 \times 40 \times 10}{40 + 10} + 1 = 17$$

$$\sigma_R = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)} = \sqrt{\frac{2 \times 40 \times 10 (2 \times 40 \times 10 - 40 - 10)}{(40 + 10)^2 (40 + 10 - 1)}} = 2.213$$

$$z = \frac{13 - 27}{2.213} = -1.81$$

Since $z_{\text{cal}} (= -1.81)$ is greater than $z_{\alpha/2} = -1.96$, reject the null hypothesis.

15.24 H_0 : There is no difference in perception.

H_1 : There is a difference in perception.

Member	Before	After	Difference	Rank
A	8	9	+ 1	2
B	3	4	+ 1	2
C	6	4	- 2	4
D	4	1	- 3	5.5
E	5	6	+ 1	2
F	7	7	0	-
G	6	9	+ 3	5.5
H	7	2	- 5	7

Sum of ranks: $s_+ = 11.5; s_- = 16.5$; Consider the smaller test statistic, $s_+ = 11.5$, which is less than its critical value, accept null hypothesis.

15.25 H_0 : Both samples come from the same population.

H_1 : The two samples come from different populations.

Branch 1	Order	Branch 2	Order
23,500	5	24,000	7.5
25500	11	19,800	2
35,500	14	22,000	4
19,500	1	21,500	3
24,400	9	24,500	10
24,000	7.5		
23,600	6		
25,900	12		
26,000	13		

Rank sum are $R_1 = 78.5$ and $R_2 = 26.5$

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

$$= 9 \times 5 + \frac{9 \times 6}{2} - 78.5 = -6.5$$

$$U_2 = 45 + 27 - 26.5 = 45.5$$

$$\mu = \frac{n_1 n_2}{2} = \frac{54}{2} = 27$$

$$\sigma_u = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{45 \times 15}{12}} = 7.5$$

$$z = \frac{U - \mu}{\sigma_u} = \frac{-6.5 - 27.0}{7.5} = -4.466$$

Since $z_{\text{cal}} < z_{\alpha}$ ($= 1.645$) accept H_0 .

$$15.26 \quad z = \frac{T - \mu_T}{\sigma_T} = \frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

$$= \frac{163.5 - (25)(26)/4}{\sqrt{(25)(26)(51)/24}} = 0.027$$

Since z_{cal} ($= 0.027$) $< z_{\alpha}$ (1.645), H_0 is accepted.

15.30 Let $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 \neq \mu_2$

Family Pair	Delhi	Mumbai	d	Rank
1	1950	1760	+ 190	+ 4
2	1840	1870	- 30	- 1
3	2015	1810	+ 205	+ 5
4	1580	1660	- 80	- 2
5	1790	1340	+ 450	+ 6
6	1925	1765	+ 160	+ 3

Since $s = \min(s_+, s_-) = \min(18, 3) = 3$ is greater than its critical value $s = 1$, H_0 is accepted.

Formulae Used

1. χ^2 -test statistic $\chi^2 = \sum(O - E)^2/E$

2. Expected frequencies for a contingency table

$$E_{ij} = \frac{\text{Row } i \text{ total} \times \text{Column } j \text{ total}}{\text{Sample size}}$$

3. Degree of freedom for a contingency table

$$df = (r - 1)(c - 1)$$

Review Self-Practice Problems

- 15.31** 1000 students at college level were graded according to their IQ and the economic conditions of their homes. Use the χ^2 -test to find out whether there is any association between economic condition at home and IQ level.

Economic Condition	IQ Level			Total
	High	Low	Total	
Rich	460	140	600	
Poor	240	160	400	
Total	700	300	1000	

[Osmania, MBA, 1998; Kumaon; Univ., MBA, 1999]

- 15.32** An automobile company gives you the following information about age groups and the liking for a particular model of car which it plans to introduce.

Opinion	Age Groups				Total
	Below 20	20-39	40-59	60 and above	
Liked the car	140	80	40	20	280
Disliked the car	60	50	30	80	220
Total	200	130	70	100	500

On the basis of this data can it be concluded that the model's appeal is independent of the age group?

[HP Univ., MCom; MD Univ., MCom, 1996]

- 15.33** The following results were obtained when two sets of items were subjected to two different treatments X and Y, to enhance their tensile strength.

- Treatment X was applied on 400 items and 80 were found to have gained in strength.
- Treatment Y was applied on 400 items and 20 were found to have gained in strength.
- Is treatment Y superior to treatment X?

[Calcutta Univ., MCom, 1998; Madras Univ., MCom, 1996]

- 15.34** Three samples are taken comprising 120 doctors, 150 advocates, and 130 university teachers. Each person chosen is asked to select one of the three categories that best represents his feeling towards a certain national policy. The three categories are in favour of policy (F), against the policy (A), and indifferent toward policy (I). The results of the interviews are given below:

Occupation	Reaction			Total
	F	A	I	
Doctors	80	30	10	120
Advocates	70	40	40	150
University teachers	50	50	30	130
Total	200	120	80	400

On the basis of this data can it be concluded that the views of doctors, advocates, and university teachers are homogeneous insofar as the National Policy under discussion is concerned?

- 15.35** Following information is obtained in a sample survey:

Condition of Child	Condition of Home		Total
	Clean	Dirty	
Clean	70	50	120
Fairly clean	80	20	100
Dirty	35	45	80
Total	185	115	300

State whether the two attributes, that is, condition of home and condition of child are independent. Use the χ^2 -test for the purpose.

[Madurai Kamraj Univ., MCom, 1996]

- 15.36** National Healthcare Company samples its hospital employees' attitude towards performance. Respondents are given a choice between the present method of two reviews a year and a proposed new method of quality reviews. The responses are given below:

	North	South	East	West
Method I	68	75	79	57
Method II	32	45	31	33

Test whether there is any significant difference in the attitude of employees in different regions at 5 per cent level of significance. [Bharthidasam Univ., MCom, 1996]

- 15.37** A milk producers union wishes to test whether the preference pattern of consumers for its products is

dependent on income levels. A random sample of 500 individuals gives the following data:

Income	Product Preferred		
	A	B	C
Low	170	30	80
Medium	50	25	60
High	20	10	55

Can you conclude that the preference patterns are independent of income levels?

[Calicut Univ., MCom, 1999]

- 15.38** The following data relate to the sales, in a time of trade depression of a certain article in wide demand. Do the data suggest that the sales are significantly affected by depression?

Pattern of Sales	Districts		
	Not Hit by Depression	Hit	Total
Satisfactory	140	60	200
Not satisfactory	40	60	100
Total	180	120	300

[GND Univ., MCom, 1997]

- 15.39** A controlled experiment was conducted to test the effectiveness of a new drug. Under this experiment 300 patients were treated with new drug and 200 were not treated with the drug. The results of the experiment are given below:

Details	Cured	Condition Worsened	No.	Total
			Effect	
Treated with the drug	200	40	60	300
Not treated with the drug	120	30	50	200
Total	320	70	110	500

Use the χ^2 -test and comment on the effectiveness of the drug.

- 15.40** 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?

- 15.41** The following figures show the distribution of digits in numbers chosen at random from a telephone directory:

Digit	:	0	1	2	3	4	5
Frequency	:	1026	1107	997	966	1075	933
Digit	:	6	7	8	9		
Frequency	:	1107	972	964	853	= 10,000	

Test whether the digits may be taken to occur equally frequently in the directory. [Roorkee Univ., MBA, 2000]

- 15.42** The number of customers that arrived in 128, 5-minute time periods at a service window were recorded as:

Customer	:	0	1	2	3	4	5
Frequency	:	2	8	10	12	18	22
Customer	:	6	7	8	9		
Frequency	:	22	16	12	6		

Is the probability distribution for the customer arrivals a Poisson distribution with a 0.05 level of significance?

- 15.43** A production supervisor is interested in knowing if the number of breakdowns of four machines is independent of the shift using the machines. Test this hypothesis based on the following sample information:

Shift	Machine				Total
	A	B	C	D	
Morning	15	10	18	12	55
Evening	12	8	15	10	45
Total	27	18	33	22	100

- 15.44** You are given the distribution of the number of defective units produced in a single shift in a factory over 100 shifts.

Number of defective units	:	0	1	2	3	4	5	6
Number of shifts:	4	14	23	23	18	9	9	

Would you say that the defective units follow a Poisson distribution? [Delhi Univ., MBA, 1995]

- 15.45** The number of car accidents that occurred during the various days of the week are as follows:

Day	No. of accidents
Sun.	14
Mon.	16
Tues.	8
Wed.	12
Thurs.	11
Fri.	9
Sat.	14

Find whether the accidents are uniformly distributed over the week. [MD Univ., MCom, 1999]

Hints and Answers

- 15.31** Let H_0 : No association between economic condition and IQ level.

O	E	$(O - E)^2$	$(O - E)^2/E$
460	420	1600	3.810
240	280	1600	5.714
140	180	1600	8.889
160	120	1600	<u>13.333</u>
			<u>31.746</u>

Since $\chi_{\text{cal}}^2 = 31.746$ is more than its critical value $\chi^2 = 3.84$ for $df = (2 - 1)(2 - 1) = 1$ and $\alpha = 5$ per cent, the H_0 is rejected.

- 15.32** Let H_0 : Personal liking is independent of the age group.

O	E	$(O - E)^2$	$(O - E)^2/E$
140	112.0	784.00	7.000
60	88.0	784.00	8.910
80	72.8	51.84	0.712
50	57.2	51.84	0.906
40	39.2	0.64	0.016
30	30.8	0.64	0.021
20	56.0	1296.00	23.143
80	44.0	1296.00	<u>29.454</u>
			<u>70.162</u>

Since $\chi_{\text{cal}}^2 = 7.162$ is more than its critical value $\chi^2 = 7.815$ for $df = (2 - 1)(4 - 1) = 3$ and $\alpha = 5$ per cent, the H_0 is rejected.

- 15.33** Let H_0 : No difference in the types of treatments X and Y.

Treatment	Gained	Not Gained	Total
X	80	320	400
Y	<u>20</u>	<u>380</u>	<u>400</u>
Total	100	700	800
O	E	$(O - E)^2$	$(O - E)^2/E$
80	50	900	18.000
20	50	900	18.000
320	350	900	2.571
380	350	900	<u>2.571</u>
			<u>41.142</u>

Since $\chi_{\text{cal}}^2 = 41.142$ is more than its critical value $\chi^2 = 3.84$ for $df = (2 - 1)(2 - 1) = 1$ and $\alpha = 5$ per cent, the H_0 is rejected.

- 15.34** Let H_0 : Opinion of three different groups of people about National Policy is same.

$$E_{11} = \frac{120 \times 200}{400} = 60 \quad E_{12} = \frac{120 \times 120}{400} = 36$$

$$E_{21} = \frac{150 \times 200}{400} = 75 \quad E_{22} = \frac{150 \times 120}{400} = 45$$

O	E	$(O - E)^2$	$(O - E)^2/E$
80	60	400	6.667
70	75	25	0.333
50	65	225	3.462
30	36	36	1.000
40	45	25	0.556
50	39	121	3.103
10	24	196	8.167
40	30	100	3.333
30	26	16	<u>0.616</u>
			<u>27.237</u>

Since $\chi_{\text{cal}}^2 = 27.237$ is more than its critical value $\chi^2 = 9.488$ for $df = (3 - 1)(3 - 1) = 4$ and $\alpha = 5$ per cent, the H_0 is rejected.

- 15.35** Let H_0 : Two attributes, that is, condition of home and condition of child are independent.

O	E	$(O - E)^2$	$(O - E)^2/E$
70	74	16	0.216
80	62	324	5.226
35	49	196	4.000
50	46	16	0.348
20	38	324	8.526
45	31	196	<u>6.322</u>
			<u>24.638</u>

Since $\chi_{\text{cal}}^2 = 24.638$ is more than its critical value $\chi^2 = 5.99$ for $df = (3 - 1)(2 - 1) = 2$ and $\alpha = 5$ per cent, the H_0 is rejected.

- 15.36** Let H_0 : No difference in the attitude of employees in different regions

O	E	$(O - E)^2$	$(O - E)^2/E$
68	66	4	0.061
32	34	4	0.118
75	80	25	0.312
45	40	25	0.625
79	73	36	0.493
31	37	36	0.973
57	60	9	0.150
33	30	9	<u>0.300</u>
			<u>3.032</u>

Since $\chi_{\text{cal}}^2 = 3.032$ is less than its critical value $\chi^2 = 7.82$ at $df = (2 - 1)(4 - 1) = 3$ and $\alpha = 5$ per cent, the H_0 is accepted.

- 15.37** Let H_0 : Income level and product preferred are independent.

O	E	$(O - E)^2$	$(O - E)^2/E$
170	134.40	1267.36	9.430
50	64.80	219.04	3.380
20	40.80	432.64	10.604
30	36.40	40.96	1.125
25	17.55	55.50	3.162
10	11.05	1.10	0.099
80	109.20	852.64	7.808
60	52.65	54.02	1.026
55	33.15	477.42	14.402
			51.036

Since $\chi^2_{\text{cal}} = 51.036$ is more than its critical value $\chi^2 = 14.860$ at $df = (3 - 1)(3 - 1) = 4$ and $\alpha = 5$ per cent, the H_0 is rejected.

- 15.38** Let H_0 : Sales are not significantly affected by depression.

O	E	$(O - E)^2$	$(O - E)^2/E$
140	120	400	3.333
40	60	400	6.667
60	80	400	5.000
60	40	400	10.000
			25.000

Since $\chi^2_{\text{cal}} = 25$ is more than its critical value $\chi^2 = 3.84$ at $df = (2 - 1)(2 - 1) = 1$ and $\alpha = 5$ per cent, the H_0 is rejected.

- 15.39** Let H_0 : No significant difference in patients, condition between those treated with the new drug and those not treated

O	E	$(O - E)^2$	$(O - E)^2/E$
200	192	64	0.333
120	128	64	0.500
40	42	4	0.095
30	28	4	0.143
60	66	36	0.545
50	44	36	0.818
			2.434

Since $\chi^2_{\text{cal}} = 2.434$ is less than its critical value $\chi^2 = 5.49$ at $df = (2 - 1)(3 - 1) = 2$ and $\alpha = 5$ per cent, the H_0 is accepted.

- 15.40** Let H_0 : Male and female births are equally probable.

Given $p = q = 0.5$. Since events are only two and mutually exclusive, the binomial distribution is used to calculate the expected frequencies (number of families). Thus the expected number of families having x male in a family of 5 children is given by

$$nP(x = r) = n^n C_r p^r q^{n-r} = 320^5 C_r (0.5)^5, r = 0, 1, 2, \dots, 5$$

O-values:	14	56	110	88	40	12
E-values:	10	50	100	100	50	10

Since $\chi^2_{\text{cal}} = \Sigma(O - E)^2/E = 7.16$ is less than its critical value $\chi^2 = 11.07$ at $df = n - 1 = 6 - 1 = 5$ and $\alpha = 5$ per cent, the null hypothesis is accepted.

- 15.41** Let H_0 : Digits are uniformly distributed in the directory.

The expected frequency (E) for each digit 0 to 9 is $10,000/10 = 1000$.

Since $\chi^2_{\text{cal}} = 58.542$ is more than its critical value $\chi^2 = 16.919$ at $df = 10 - 1 = 9$ and $\alpha = 5$ per cent, the null hypothesis is rejected.

- 15.42** Let H_0 : The population has a Poisson distribution

Total number of customers who arrived during the sample of 128, 5-minute time periods is given by $0(2) + 1(8) + 2(10) + \dots + 9(6) = 640$. Thus 640 customers arrival over a sample of 128 periods provide a mean arrival rate of $\lambda = 640/128 = 5$ customers per 5-minute period.

The expected frequency (E) of customer arrivals for each of the variable from 0 to 9 is calculated by using the Poisson distribution formula

$$nP(x = r) = n \frac{e^{-\lambda} \lambda^r}{r!} = 128 \frac{e^{-5} (5)^r}{r!}$$

$$r = 0, 1, \dots, 9; \lambda = 5$$

Number of customers	Observed frequency	Expected frequency
0 or 1	10	5.17
2	10	10.77
3	12	17.97
4	18	22.46
5	22	22.46
6	22	18.71
7	16	13.36
8	12	8.35
9 or more	6	8.71

Since $\chi^2_{\text{cal}} = 10.97$ is less than its critical value $\chi^2 = 14.07$ at $df = n - 1 - 2 = 10 - 1 - 2 = 7$ and $\alpha = 0.05$, the null hypothesis is accepted.

- 15.43** Let H_0 : Number of breakdowns is independent of the shift using the machines.

$$\text{Expected frequencies} = \frac{\text{Row } i \text{ total} \times \text{Column } j \text{ total}}{\text{Grand total}}$$

O-values :	15	12	10	8	18	15	12	10
E-values :	14.85	12.15	9.90	8.10	18.15	14.85	12.10	9.90

This page is intentionally left blank.

Appendices

LEARNING OBJECTIVES

After studying this appendix, you should be able to

- Table A 1: Poisson Probabilities
- Table A 2: Binomial Coefficients
- Table A 3: Normal Distribution
- Table A 4: Critical Values of t -Distribution
- Table A 5: Critical Values of Chi-Square (χ^2)
- Table A 6: 5 Per cent Points of Fisher's F-Distribution
- Table A 7: Critical Values for Sign Test
- Table A 8: Critical Values for Wilcoxon Signed-Rank Test
- Table A 9: Critical Values for Wilcoxon Rank-Sum Test
- Table A 10: Critical Values of R for Runs Test
- Table A 11: Factors for Construction of Control Charts

Table A1: Poission Probabilities

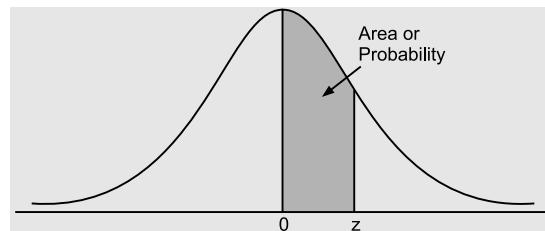
x	λ									
	0.005	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	.9950	.9900	.9802	.9704	.9608	.9512	.9418	.9324	.9231	.9139
1	.0050	.0099	.0196	.0291	.0384	.0476	.0565	.0653	.0738	.0823
2	.0000	.0000	.0002	.0004	.0008	.0012	.0017	.0023	.0030	.0037
3	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	.9048	.8187	.7408	.6703	.6065	.5488	.4966	.4493	.4066	.3679
1	.0905	.1637	.2222	.2681	.3033	.3293	.3476	.3595	.3659	.3679
2	.0045	.0164	.0333	.0536	.0758	.0988	.1217	.1438	.1647	.1839
3	.0002	.0011	.0033	.0072	.0126	.0198	.0284	.0383	.0494	.0613
4	.0000	.0001	.0003	.0007	.0016	.0030	.0050	.0077	.0111	.0153
5	.0000	.0000	.0000	.0001	.0002	.0004	.0007	.0012	.0020	.0031
6	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0005
7	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

Contd...

<i>x</i>	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	.3329	.3012	.2725	.2466	.2231	.2019	.1827	.1653	.1496	.1353
1	.3662	.3614	.3543	.3452	.3347	.3230	.3106	.2975	.2842	.2707
2	.2014	.2169	.2303	.2417	.2510	.2584	.2640	.2678	.2700	.2707
3	.0738	.0867	.0998	.1128	.1155	.1378	.1496	.1607	.1710	.1804
4	.0203	.0260	.0324	.0395	.0471	.0551	.0636	.0723	.0812	.0902
5	.0045	.0062	.0084	.0111	.0141	.0176	.0216	.0260	.0309	.0361
6	.0008	.0012	.0018	.0026	.0035	.0047	.0061	.0078	.0098	.0120
7	.0001	.0002	.0003	.0005	.0008	.0011	.0015	.0020	.0027	.0034
8	.0000	.0000	.0001	.0001	.0001	.0002	.0003	.0005	.0006	.0009
9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002
<i>x</i>	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
0	.1225	.1108	.1003	.0907	.0821	.0743	.0672	.0608	.0550	.0498
1	.2572	.2438	.2306	.2177	.2052	.1931	.1815	.1703	.1596	.1496
2	.2700	.2681	.2652	.2613	.2565	.2510	.2450	.2384	.2314	.2240
3	.1890	.1966	.2033	.2090	.2138	.2176	.2205	.2225	.2237	.2240
4	.0992	.1082	.1169	.1254	.1336	.1414	.1488	.1557	.1622	.1680
5	.0417	.0476	.0538	.0602	.0668	.0735	.0804	.0872	.0940	.1008
6	.0146	.0174	.0206	.0241	.0278	.0319	.0362	.0407	.0455	.0504
7	.0044	.0055	.0068	.0083	.0099	.0118	.0139	.0163	.0188	.0216
8	.0011	.0015	.0019	.0025	.0031	.0038	.0047	.0057	.0068	.0081
9	.0003	.0004	.0005	.0007	.0009	.0011	.0014	.0018	.0022	.0027
10	.0001	.0001	.0001	.0002	.0002	.0003	.0004	.0005	.0006	.0008
11	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002	.0002
<i>x</i>	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	.0450	.0408	.0369	.0334	.0302	.0273	.0247	.0224	.0202	.0183
1	.1397	.1304	.1217	.1135	.1057	.0984	.0915	.0850	.0789	.0733
2	.2165	.2087	.2008	.1929	.1850	.1771	.1692	.1615	.1539	.1459
3	.2237	.2226	.2209	.2186	.2158	.2125	.2087	.2046	.2001	.1954
4	.1733	.1781	.1823	.1858	.1888	.1912	.1931	.1944	.1951	.1954
5	.1075	.1140	.1203	.1265	.1322	.1377	.1429	.1477	.1522	.1563
6	.0555	.0608	.0662	.0716	.0771	.0826	.0881	.0936	.0989	.1042
7	.0246	.0278	.0312	.0348	.0385	.0425	.0466	.0508	.0551	.0595
8	.0095	.0111	.0129	.0148	.0169	.0191	.0225	.0241	.0269	.0298
9	.0033	.0040	.0047	.0056	.0066	.0076	.0089	.0102	.0116	.0132
10	.0010	.0013	.0016	.0019	.0023	.0028	.0033	.0039	.0045	.0053
11	.0003	.0004	.0005	.0006	.0007	.0009	.0011	.0013	.0016	.0019
12	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005	.0006
13	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002	.0002	.0002
14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
<i>x</i>	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	.0166	.0150	.0136	.0123	.0111	.0101	.0191	.0082	.0074	.0067
1	.0679	.0630	.0583	.0540	.0500	.0462	.0427	.0395	.0365	.0337
2	.1393	.1323	.1254	.1188	.1125	.1063	.1005	.0948	.0894	.0842
3	.1904	.1852	.1798	.1743	.1687	.1631	.1574	.1517	.1460	.1404
4	.1951	.1944	.1933	.1917	.1898	.1875	.1849	.1820	.1789	.1755
5	.1600	.1633	.1662	.1687	.1708	.1725	.1738	.1747	.1753	.1755
6	.1093	.1143	.1191	.1237	.1281	.1323	.1362	.1398	.1432	.1462
7	.0640	.0686	.0732	.0778	.0824	.0869	.0914	.0959	.1002	.1044
8	.0328	.0360	.0393	.0428	.0463	.0500	.0537	.0575	.0614	.0653
9	.0150	.0168	.0188	.0209	.0232	.0255	.0281	.0307	.0334	.0363
10	.0061	.0071	.0081	.0092	.0104	.0118	.0132	.0147	.0164	.0181
11	.0023	.0027	.0032	.0037	.0043	.0049	.0056	.0064	.0073	.0082
12	.0008	.0009	.0011	.0013	.0016	.0019	.0022	.0026	.0030	.0034
13	.0002	.0003	.0004	.0005	.0006	.0007	.0008	.0009	.0011	.0013
14	.0001	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005
15	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0002

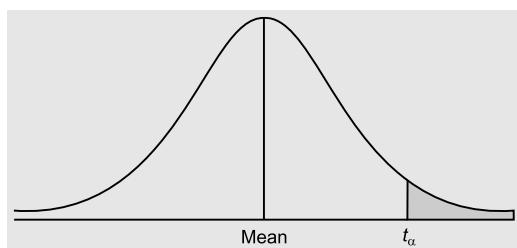
Table A2: Binomial Coefficients

n	nC_0	nC_1	nC_2	nC_3	nC_4	nC_5	nC_6	nC_7	nC_8	nC_9	${}^nC_{10}$
0	1										
1	1	1									
2	1	2	1								
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1
11	1	11	55	165	330	462	462	330	165	55	11
12	1	12	66	220	495	792	924	792	495	220	66
13	1	13	78	286	715	1287	1716	1716	1287	715	286
14	1	14	91	364	1001	2002	3003	3432	3003	2002	1001
15	1	15	105	455	1365	3003	5005	6435	6435	3005	3003
16	1	16	120	560	1820	4368	8008	11440	12870	11440	8008
17	1	17	136	680	2380	6188	12376	19448	24310	24310	19448
18	1	18	153	816	3060	8568	18564	31824	43758	48620	43758
19	1	19	171	969	3876	11628	27132	50388	75582	92378	92378
20	1	20	190	1140	4845	15504	38760	77520	125970167960	184756	

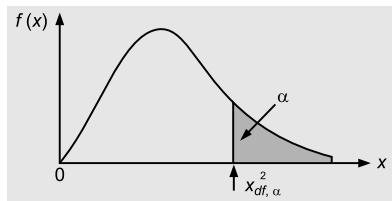
Table A3: Area of Standard Normal Distribution

Areas under the standard normal probability distribution between normal variate $z = 0$ and a positive value of z . Areas for negative value of z are obtained by symmetry.

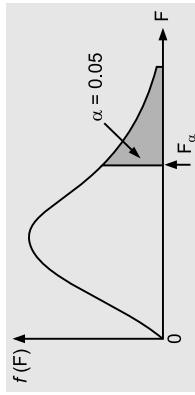
Value of z First decimal place x	Second Decimal									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2611	.2642	.2674	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4865	.4868	.4871	.4874	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4986	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Table A4: Critical Values of Student's *t*-Distribution

Level of Significance for One-tailed Test					
<i>d.f.</i>	0.10	0.05	0.025	0.01	0.005
Level of Significance for One-tailed Test					
<i>d.f.</i>	0.20	0.10	0.05	0.02	0.01
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750

Table A5: Critical Values of Chi-Square (χ^2)

Degree of freedom	0.100	0.050	0.025	0.010	0.005	0.001
1	2.71	3.84	5.02	6.63	7.88	10.8
2	4.61	5.99	7.38	9.21	10.6	13.8
3	6.25	7.81	9.35	11.3	12.8	16.3
4	7.78	9.49	11.1	13.3	14.9	18.5
5	9.24	11.1	12.8	15.1	16.7	20.5
6	10.6	12.6	14.4	16.8	18.5	22.5
7	12.0	14.1	16.0	18.5	20.3	24.3
8	13.4	15.5	17.5	20.1	22.0	26.1
9	14.7	16.9	19.0	21.7	23.6	27.9
10	16.0	18.3	20.5	23.2	25.2	29.6
11	17.3	19.7	21.9	24.7	26.8	31.3
12	18.5	21.0	23.3	26.2	28.3	32.9
13	19.8	22.4	24.7	27.7	29.8	34.5
14	21.1	23.7	26.1	29.1	31.3	36.1
15	22.3	25.0	27.5	30.6	32.8	37.7
16	23.5	26.3	28.8	32.0	34.3	39.3
17	24.8	27.6	30.2	33.4	35.7	40.8
18	26.0	28.9	31.5	34.8	37.2	42.3
19	27.2	30.1	32.9	36.2	38.6	43.8
20	28.4	31.4	34.2	37.6	40.0	45.3
21	29.6	32.7	35.5	38.9	41.4	46.8
22	30.8	33.9	36.8	40.3	42.8	48.3
23	32.0	35.2	38.1	41.6	44.2	49.7
24	33.2	36.4	39.4	43.0	45.6	51.2
25	34.4	37.7	40.6	44.3	46.9	52.6
26	35.6	38.9	41.9	45.6	48.3	54.1
27	36.7	40.1	43.2	47.0	49.6	55.5
28	37.9	41.3	44.5	48.3	51.0	56.9
29	39.1	42.6	45.7	49.6	52.3	58.3
30	40.3	43.8	47.0	50.9	53.7	59.7
35	46.1	49.8	53.2	57.3	60.3	66.6
40	51.8	55.8	59.3	63.7	66.8	73.4
45	57.5	61.7	65.4	70.0	73.2	80.1
50	63.2	67.5	71.4	76.2	79.5	86.7

Table A6: Critical Values of the F-Distribution at a 5 Per cent Level of Significance

Degrees of freedom for the denominator	Degrees of freedom for the Numerator									
	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8868	8.8552	8.8123	8.7855
4	7.7086	6.9443	6.5914	6.3883	6.2560	6.1631	6.0942	6.0410	5.9644	5.9117
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8753	4.8183	4.7725	4.7351
6	5.9875	5.1433	4.7571	4.5337	4.3874	4.2839	4.2066	4.1468	4.0990	4.0600
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365
8	5.3177	4.4590	4.0662	3.8378	3.6875	3.5806	3.5005	3.4381	3.3472	3.2840
9	5.1174	4.29565	3.8626	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536
12	4.7272	3.8853	3.4903	3.2502	3.1059	2.9961	2.9134	2.8486	2.7534	2.6866
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6021
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377	2.4935
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943	2.4499
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563	2.4117
19	4.3808	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	2.3779
20	4.3513	3.4928	3.0984	2.8661	2.7100	2.5990	2.5140	2.4471	2.3928	2.3479
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3661	2.3210
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419	2.2967
23	4.2793	3.4221	3.0280	2.7955	2.6500	2.5277	2.4422	2.3748	2.3201	2.2747
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002	2.2547
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	2.2365
26	4.2252	3.3690	2.9751	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655	2.2197
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501	2.2043
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360	2.1900
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2782	2.2229	2.1768
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646
31	4.0848	3.2317	2.8387	2.6037	2.4495	2.3359	2.2490	2.1802	2.1240	2.0772
32	4.0012	3.1504	2.7581	2.5252	2.3688	2.2540	2.1665	2.0970	2.0401	1.9926
33	3.9201	3.0718	2.6802	2.4472	2.2900	2.1750	2.0867	2.0164	1.9588	1.9105
34	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799	1.8307

Table A7: Critical Values for Sign Test

$\begin{array}{c} \diagup \\ n \end{array}$	a	0.10 0.05	0.05 0.025	0.01 0.005	Two-sided One-sided tests	$\begin{array}{c} \diagup \\ n \end{array}$	a	0.10 0.05	0.05 0.025	0.01 0.005	Two-sided tests One-sided tests
5	0					23	7	6	4		
6	0	0				24	7	6	5		
7	0	0				25	7	7	5		
8	1	0	0			26	8	7	6		
9	1	1	0			27	8	7	6		
10	1	1	0			28	9	8	6		
11	2	1	0			29	9	8	7		
12	2	2	1			30	10	9	7		
13	3	2	1			31	10	9	7		
14	3	2	1			32	10	9	8		
15	3	3	2			33	11	10	8		
16	4	3	2			34	11	10	9		
17	4	4	2			35	12	11	9		
18	5	4	3			36	12	11	9		
19	5	4	3			37	13	12	10		
20	5	5	3			38	13	12	10		
21	6	5	4			39	13	12	11		
22	6	5	4			40	14	13	11		

Table A.8: Critical Values for the Wilcoxon Signed-Rank Test

$\begin{array}{c} \diagup \\ n^* \end{array}$	a	0.10 0.05	0.05 0.025	0.01 0.005	Two-sided One-sided tests
4					
5	0				
6	2	0			
7	3	2	0		
8	5	3	1	0	
9	8	5	3	1	
10	10	8	5	3	
11	13	10	7	5	
12	17	13	9	7	
13	21	17	12	9	
14	25	21	15	12	
15	30	25	19	15	
16	35	29	23	19	
17	41	34	27	23	
18	47	40	32	27	
19	53	46	37	32	
20	60	52	43	37	
21	67	58	49	42	
22	75	65	55	48	
23	83	73	62	54	
24	91	81	69	61	
25	100	89	76	68	

* If $n > 25$, W^+ (or W^-) is approximately normally distributed with mean $n(n + 1)/4$ and variance $n(n + 1)(2n + 1)/24$.

Table A.9: Critical Values for the Wilcoxon Rank-Sum Test ($\lambda = 0.05$)

* For $n_1 > 8$, W_1 is approximately normally distributed with mean $1/2 n_1(n_1 + n_2 + 1)$ and variance $n_1 n_2 (n_1 + n_2 + 1)/12$.

Critical Values for the Wilcoxon Rank-Sum Test ($\lambda = 0.01$)

Table A.10: Critical Values of R for the Runs Test (Lower Tail)

$n_2 \backslash n_1$	$\alpha = 0.026$																			
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
2											2	2	2	2	2	2	2	2	2	2
3					2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	
4				2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	
5			2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5	
6		2	2	3	3	3	3	4	4	4	4	4	5	5	5	5	5	6	6	
7		2	2	3	3	3	4	4	5	5	5	5	5	6	6	6	6	6	6	
8		2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7	
9		2	3	3	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	
10		2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9	
11		2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9	
12		2	2	3	4	4	5	6	6	7	7	7	8	8	8	9	9	9	10	
13		2	2	3	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10	
14		2	2	3	4	5	5	6	7	7	8	8	9	9	9	10	10	10	11	
15		2	3	3	4	5	6	6	7	7	8	8	9	9	10	10	11	11	12	
16		2	3	4	4	5	6	6	7	8	8	9	9	10	10	11	11	11	12	
17		2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12	
18		2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12	13	
19		2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	
20		2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	14	

Critical Values of R for the Runs Test (Upper Tail)

$n_2 \backslash n_1$	$\alpha = 0.025$																			
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
2																				
3																				
4			9	9																
5			9	10	10	11	11													
6			9	10	11	12	12	13	13	13	13									
7			11	12	13	13	14	14	14	14	14	15	15	15	15					
8			11	12	13	14	14	15	15	15	16	16	16	16	16	17	17	17	17	17
9			13	14	14	15	16	16	16	16	17	17	18	18	18	18	18	18	18	18
10			13	14	15	16	16	17	17	17	18	18	18	18	19	19	19	19	20	20
11			13	14	15	16	17	17	17	18	19	19	19	19	20	20	20	20	21	21
12			13	14	16	16	17	18	19	19	19	20	20	20	21	21	21	22	22	22
13			15	16	17	18	19	19	19	20	20	21	21	21	22	22	23	23	23	23
14			15	16	17	18	19	20	20	20	21	21	22	22	22	23	23	23	24	24
15			15	16	18	18	19	20	20	21	22	22	22	23	23	24	24	25	25	25
16				17	18	19	20	21	21	22	22	23	23	24	24	25	25	25	25	25
17				17	18	19	20	21	22	22	23	23	24	24	25	25	26	26	26	26
18				17	18	19	20	21	22	22	23	23	24	24	25	25	26	26	27	27
19				17	18	20	21	22	23	23	24	24	25	25	26	26	26	27	27	27
20				17	18	20	21	22	23	24	25	25	26	26	27	27	27	28		

Table A.11: *p*-Values for Mann-Whitney U Statistic Small Samples ($n_1 \leq n_2$)

		n_1		
$n_2 = 3$	U_0	1	2	3
	0	.25	.10	.05
	1	.50	.20	.10
	2		.40	.20
	3		.60	.35
	4			.50

		n_1			
$n_2 = 4$	U_0	1	2	3	4
	0	.2000	.0667	.0286	.0143
	1	.4000	.1333	.0571	.0286
	2	.6000	.2667	.1143	.0571
	3		.4000	.2000	.1000
	4		.6000	.3143	.1714
	5			.4286	.2429
	6			.5714	.3429
	7				.4429
	8				.5571

		n_1				
$n_2 = 5$	U_0	1	2	3	4	5
	0	.1667	.0476	.0179	.0079	.0040
	1	.3333	.0952	.0357	.0159	.0079
	2	.5000	.1905	.0714	.0317	.0159
	3		.2857	.1250	.0556	.0278
	4		.4286	.1964	.0952	.0476
	5		.5714	.2857	.1429	.0754
	6			.3929	.2063	.1111
	7			.5000	.2778	.1548
	8				.3651	.2103
	9				.4524	.2738
	10				.5476	.3452
	11					.4206
	12					.5000

		n_1					
$n_2 = 6$	U_0	1	2	3	4	5	6
	0	.1429	.0357	.0119	.0048	.0022	.0011
	1	.2857	.0714	.0238	.0095	.0043	.0022
	2	.4286	.1429	.0476	.0190	.0087	.0043
	3	.5714	.2143	.0833	.0333	.0152	.0076
	4		.3214	.1310	.0571	.0260	.0130
	5		.4286	.1905	.0857	.0411	.0206
	6		.5714	.2738	.1286	.0628	.0325
	7			.3571	.1762	.0887	.0465
	8				.4524	.2381	.1234
	9				.5476	.3048	.1645
	10					.3810	.2143
	11					.4571	.2684
	12					.5429	.3312
	13						.3961
	14						.4654
	15						.5346
	16						.4091
	17						.4686
	18						.5314

Table A. 12: Factors Useful in the Construction of Control Charts

Sample Size	Mean-Chart			Factors for Central Line				Factors for Control Limit				Range-Chart			
	<i>A</i>	<i>A</i> ₁	<i>A</i> ₂	<i>c</i> ₂	<i>B</i> ₁	<i>B</i> ₂	<i>B</i> ₃	<i>B</i> ₄	<i>d</i> ₂	<i>D</i> ₁	<i>D</i> ₂	<i>D</i> ₃	<i>D</i> ₄	Factors for Central Line	Factors for Control Limit
2	2.121	3.760	1.881	0.6642	0	1.843	0	3.267	1.128	0	3.686	0	3.267		
3	1.732	3.394	1.023	0.7236	0	1.858	0	2.566	1.693	0	4.358	0	2.575		
4	1.500	2.880	0.729	0.7979	0	1.808	0	2.269	2.059	0	4.698	0	2.282		
5	1.342	1.596	0.577	0.8407	0	1.756	0	2.089	2.326	0	4.918	0	2.115		
6	1.225	1.410	0.483	0.8686	0.026	1.711	0.030	1.970	2.534	0	5.078	0	2.004		
7	1.134	1.277	0.419	0.8882	0.105	1.672	0.118	1.888	2.704	2.205	5.203	0.076	1.924		
8	1.061	1.175	0.073	0.9027	0.167	1.638	0.185	1.815	2.847	0.387	5.307	0.136	1.864		
9	1.000	1.094	0.037	0.9139	0.219	1.609	0.239	1.761	2.970	0.546	5.394	0.184	1.816		
10	0.949	1.028	0.308	0.9227	0.262	1.584	0.284	1.716	3.078	0.687	5.469	0.223	1.777		
11	0.905	0.973	0.285	0.9300	0.299	1.561	0.321	1.679	3.173	0.812	5.534	0.256	1.744		
12	0.866	0.925	0.256	0.9359	0.331	1.541	0.354	1.646	3.258	0.924	5.592	0.284	1.716		
13	0.832	0.883	0.249	0.9410	0.359	1.523	0.382	1.618	3.336	1.026	5.646	0.308	1.692		
14	0.802	0.848	0.235	0.9453	0.384	1.507	0.406	1.594	3.407	1.121	5.693	0.329	1.671		
15	0.775	0.816	0.223	0.9490	0.406	1.492	0.428	1.572	3.472	1.207	5.737	0.348	1.652		
16	0.750	0.788	0.212	0.9523	0.427	1.478	0.448	1.542	3.352	1.285	5.279	0.365	1.636		
17	0.728	0.762	0.203	0.9551	0.445	1.465	0.466	1.534	3.588	1.359	5.817	0.379	1.621		
18	0.707	0.738	0.816	0.9576	0.461	1.454	0.482	1.518	3.640	1.426	5.854	0.404	1.608		
19	0.688	0.717	0.187	0.9599	0.477	1.443	0.497	1.503	3.689	1.490	5.888	0.404	1.596		
20	0.671	0.697	0.180	0.9619	0.491	1.433	0.510	1.490	3.735	1.548	5.922	0.414	1.585		
21	0.655	0.670	0.173	0.9638	0.504	1.424	0.523	1.447	3.778	1.606	5.950	0.425	1.575		
22	0.640	0.662	0.167	0.9655	0.516	1.415	0.534	1.466	3.819	1.659	5.979	0.434	1.566		
23	0.626	0.647	0.162	0.9670	0.527	1.407	0.545	1.455	3.858	1.710	6.006	0.443	1.557		
24	0.612	0.632	0.157	0.9684	0.538	1.399	0.555	1.445	3.895	1.759	6.031	0.452	1.548		
25	0.600	0.319	0.153	0.9696	0.548	1.392	0.565	1.435	3.931	1.804	6.058	0.459	1.541		

This page is intentionally left blank.

MODEL QUESTION PAPER-I
MBA Degree Examination
STATISTICS FOR MANAGEMENT

Time: 3 hrs.

Max. Marks: 100

- Note:** 1. Answer any **FIVE** full questions.
 2. Use of scientific non-programmable calculators is permitted.

1. (a) What is classification? Mention the objectives and bases of classification. **(07 Marks)**
 (b) The average daily wages of all the workers in a factory is Rs. 444. If the average daily wages paid to male and female workers are Rs. 480 and Rs. 360 respectively, find the percentage of male and female workers in that factory. **(05 Marks)**
 (c) Draw a histogram of the following distribution and hence locate the mode graphically:

C.I.	0-4	4-8	8-12	12-20	20-24	24-32	32-40	Total
f	6	8	10	24	13	16	8	85

(08 Marks)

2. (a) Distinguish between: (i) Primary data and secondary data
 (ii) Census method and sampling method. **(06 Marks)**
 (b) Calculate Spearman's rank correlation coefficient between advertising cost and sales given below:

Advertising cost (Lakh Rs.)	39	65	62	90	82	75	25	98	36	78
Sales (Lakh Rs.)		47	53	58	86	62	68	60	91	51

- (c) The following data relates to the number of accidents in 30 cities. Find Karl Pearson's co-efficient of skewness: **(08 Marks)**

No. of accidents	10	11	12	13	14	15
No. of cities	2	4	10	8	5	1

3. (a) What do you mean by 'semi inter-quartile range'? Explain its merits and demerits. **(05 Marks)**
 (b) Calculate Laspeyre's, Paasche's and Fisher's index numbers using the following data:

Item	Base year		Current year	
	Price (Rs.)	Total value (Rs.)	Price (Rs.)	Total value (Rs.)
A	6	300	10	560
B	2	200	2	240
C	4	240	6	360
D	10	300	12	288
E	8	320	12	432

(07 Marks)

- (c) The frequency distribution which has only been partly reproduced is given below:

Marks	30–35	35–40	40–45	45–50	50–55	55–60	60–65
No. of students	3	5	–	18	14	–	2

If the modal marks of the 60 students is given to be 47.5, then find the missing values. **(08 Marks)**

4. (a) Explain with suitable examples:

- | | |
|-----------------------------|--------------------------------------|
| (i) Event | (ii) Random experiment |
| (iii) Sample space | (iv) Equally likely events |
| (v) Probability of an event | (vi) Addition theorem of probability |
- (07 Marks)**

- (b) The probability that a contractor will get plumbing contract is $2/3$ and the probability that he will not get the electrical contract is $5/9$. If the probability of getting atleast one of these contracts is $4/5$, what is the probability that he will get both? **(05 Marks)**

- (c) You are given the exports of electronic goods from 1990 to 1995. Fit a linear trend to the exports data and estimate the expected exports for the year 2005. **(08 Marks)**

Year	1990	1991	1992	1993	1994	1995
Exports in crore Rs.	11	16	13	18	22	20

5. (a) What is meant by 'correlation analysis'? Explain the types of correlation. **(06 Marks)**

- (b) Compute the median and 63rd percentile from the following details. **(06 Marks)**

Weight (kgs.)	0–4	5–9	10–14	15–19	20–24	25–29
No. of bags	5	7	10	8	6	4

- (c) The height of father and sons are given below:

Height of father (inches)	65	66	67	67	68	69	71	73
Height of son (inches)	67	68	64	68	72	70	69	70

- (i) Find the two lines of regression.
(ii) Estimate the expected height of son when the height of father is 67.5 inches. **(08 Marks)**

6. (a) What is time series? Explain the uses of time series. **(05 Marks)**

- (b) Calculate mean deviation and its coefficient of the following distribution using median:

Age (yrs)	Below 10	Below 20	Below 30	Below 40	Below 50	Below 60
No. of persons	8	15	24	34	47	50

(07 Marks)

- (c) Fit a Poisson distribution to the following data (Given $e^{-0.5} = 0.6065$). **(08 Marks)**

x	0	1	2	3	4
f	123	59	14	3	1

(08 Marks)

7. (a) Explain the different methods of sampling. **(06 Marks)**
 (b) The following are the B.P.'s of 8 persons before and after meditation. Can we conclude that meditation reduces B.P? (Given for 7 d.f. $t_{0.05} = 1.90$)
(07 Marks)

Person	1	2	3	4	5	6	7	8
Before meditation	92	90	86	92	88	94	90	90
After meditation	86	88	80	86	86	84	84	90

- (c) An automobile company gives you the following information about age groups and the liking for a particular model of car that it plans to launch:

Age (years)	Below 25	25–50	Above 50
Who liked the car	45	30	25
Who disliked the car	55	20	25

On the basis of the above data, can it be concluded that the model appeal is independent of the age group? (Value of X^2 for 2 d.f. is 5.991). **(07 Marks)**

8. (a) Explain the procedure for testing a hypothesis. **(06 Marks)**
 (b) Among 80 electric bulbs manufactured by process 'A', three were defective. Among 130 electric bulbs manufactured by process 'B', two were defective. Test whether proportion of defectives in the two processes differ. **(06 Marks)**
 (c) Following are the weekly sale records (in thousand rupees) of three salesmen A, B and C of a company during 13 sale calls:

A	300	400	300	500	-
B	600	300	300	400	-
C	700	300	400	600	500

Test at 5% level of significance whether the sales of the three salesmen are different. (Given $F_{0.05} = 4.10$ for $V_1 = 2$, $V_2 = 10$). **(08 Marks)**

MODEL QUESTION PAPER-II
MBA Degree Examination
STATISTICS FOR MANAGEMENT

Time: 3 hrs.

Max. Marks: 100

Note: 1. Answer any **FIVE** full questions.

2. Use of statistical tables is permitted.

1. (a) Distinguish between primary data and secondary data and discuss the various methods of collecting secondary data. **(08 Marks)**
- (b) The amounts of interest paid on each of the three different sums of money yielding 10%, 12% and 15%, simple interest per annum are equal. What is the average yield percent on the total sum invested? **(05 Marks)**
- (c) You are working as a transport manager of a software company. Kilometers recorded for a sample of hired cars during a week yielded the following data:

Kilometers covered	No. of cars	Kilometers covered	No. of cars
100–110	4	150–160	8
110–120	0	160–170	5
120–130	3	170–180	0
130–140	7	180–190	2
140–150	11	Total	40

Form a cumulative frequency and draw a cumulative frequency Ogive. Calculate Q_1 , Q_2 , Q_3 and P_{75} . Estimate graphically the number of cars which covered less than 165 km in the week.

2. (a) What is an index number? Describe briefly its applications in business and industry. **(08 Marks)**
- (b) "After settlement the average weekly wage in a factory had increased from Rs 800 to Rs 1200 and the standard deviation has increased from Rs 100 to Rs 150". Comment on the uniformity of the wages before and after the settlement. **(05 Marks)**
- (c) A factory pays workers on a piece rate basis and also bonus to each worker on the basis of individual output in each quarter. Calculate the average bonus/worker for a quarter and average output/worker. **(07 Marks)**

Output (in units)	70–74	75–79	80–84	85–89	90–94	95–99	100–104
Frequency	3	5	15	12	7	6	2
Bonus (Rs)	40	45	50	60	70	80	100

3. (a) Distinguish between correlation and regression analysis. Provide examples to indicate their uses.
- (b) The rate of increase in population of a country during the last 3 decades is 5 percent, 8 percent and 12 percent. Find the average rate of growth during last three decades. **(05 Marks)**
- (c) Calculate the Karl Pearson's coefficient of correlation between trunk height and tree diameter from the data given below: **(07 Marks)**

Trunk height	35	49	27	33	60	21	45	51
Diameter	8	9	7	6	13	7	11	12

4. (a) "Averages, Dispersion and Skewness are complementary to one another in understanding a frequency distribution" – Explain. **(08 Marks)**

- (b) A Branch Manager of a Bank notices that, over a long period of time, the number of people using an ATM on a Saturday morning is on an average, 30 people per hour. What is the probability that in a 10 minute period.
- (i) Nobody uses the machine? (ii) 3 people use the machine? **(05 Marks)**
- (c) Find the regression equation showing the capacity utilization on production.

	<i>Average</i>	<i>Standard deviation</i>
Production (in lakh units)	35.6	10.5
Capacity utilization (in %)	84.8	8.5
Correlation coefficient	$\gamma = 0.62$	

Estimate the production when capacity utilization is 70%. **(07 Marks)**

5. (a) Distinguish between trend, seasonal and cyclical variations in a time series. What effect does seasonal variability produce? **(08 Marks)**
- (b) Represent the following data by a percentage sub-divided bar diagram. **(05 Marks)**

<i>Items of expenditure</i>	<i>Family A</i>	<i>Family B</i>
	<i>Income Rs 500</i>	<i>Income Rs 300</i>
Food	150	150
Clothing	125	60
Education	25	50
Miscellaneous	190	70
Savings or Deficit	+10	-30

- (c) Compute Laspeyre's, Paasche's and Fisher's price index from the following data: **(07 Marks)**

<i>Commodities</i>	<i>1990 Base year 2000</i>			
	<i>Price</i>	<i>Quantity</i>	<i>Price</i>	<i>Quantity</i>
A	20	8	40	6
B	50	10	60	5
C	40	15	50	15
D	20	20	20	25

6. (a) Explain the important features of normal and Poisson distribution. **(08 Marks)**
- (b) How is analysis of variance helpful in solving business problems? Illustrate your answer with suitable examples. **(05 Marks)**
- (c) The sales of a company in millions of rupees for the years 1998–2005 are given below:

<i>Year</i>	<i>Sales</i>	<i>Year</i>	<i>Sales</i>
1998	550	2003	525
1999	560	2004	545
2000	555	2005	585
2001	585		
2002	540		

- (i) Find the linear trend equation
(ii) Estimate the sales for the year 1997
(iii) Find the slope of the straight line trend. **(07 Marks)**

7. (a) Enumerate the various methods of sampling. Describe two of them, mentioning the situations where each one is to be used. **(08 Marks)**
- (b) A firm has appointed a large number of dealers all over the country to sell its food products. A random sample of 25 dealers is chosen for this purpose. The sample mean is Rs 30,000 and the sample standard deviation is Rs 10,000. Construct an interval estimate with 95% confidence. **(05 Marks)**
- (c) Two salesman *A* and *B* are employed by a company. Recently it conducted a sample survey which yielded the following data:

	<i>Salesman A</i>	<i>Salesman B</i>
No. of sales	20	22
Average sales (Rs)	800	780
Standard deviation (Rs)	70	60

Is there any significant difference between the average sales of the two salesmen?

(07 Marks)

8. (a) Describe the various steps involved in testing of a hypothesis. **(08 Marks)**
- (b) A manufacturer claims that atleast 95% of requirements supplied by him confirmed to specifications. An examination of the sample of 200 pieces of equipments revealed that 18 were faulty. Test the claim of the manufacturer. **(05 Marks)**
- (c) A personnel manager is interested in trying to determine if absenteeism is greater on any one day of the week. Past records for last 1 year, show the following sample distribution:

Day of the week	Monday	Tuesday	Wednesday	Thursday	Friday
No. of absentees	66	56	54	48	75

Test whether absence is uniformly distributed over the week. **(07 Marks)**

MODEL QUESTION PAPER-III
MBA Degree Examination
STATISTICS FOR MANAGEMENT

Time: 3 hrs.

Max. Marks: 100

Note: 1. Answer any **FIVE** full questions.

2. Use of statistical tables and non-programmable scientific calculators is allowed.

1. (a) Distinguish between the primary and secondary data. Mention the methods of collecting primary data. **(06 Marks)**

- (b) Prepare a frequency distribution for the following observations taking class intervals as 15–25, 25–35, 35–45, ... and so on. Hence draw the histogram.

15	45	40	42	50	60	62	68	70	42
75	75	80	81	25	26	31	32	78	45
31	45	42	43	55	56	78	80	81	62
60	62	58	69	70	45	50	56	72	58
75	62	62	65	60	70	35	37	40	55

(06 Marks)

- (c) The following table gives the frequency distribution of the weekly wages (in '00 Rs.) of 100 workers in a factory.

Weekly wages ('00 Rs)	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60–64	Total
No. of workers	4	5	12	23	31	10	8	5	2	100

Draw the frequency polygon and less than 0 gives for the above table.

(08 Marks)

2. (a) What is meant by measure of central tendency? What are the merits and demerits of arithmetic mean? **(06 Marks)**

- (b) In the following grouped data, X are mid values of class interval and C is constant. If the arithmetic mean of the original distribution is 35.84, find its class intervals. **(06 Marks)**

$X-C$	-21	-14	-7	0	7	14	21	Total
f	2	12	19	29	20	13	5	100

- (c) The following data gives the distribution of 100 students. Calculate the most suitable average, giving the reason for your choice. Also obtain the values of quartiles, 6th decile and 70th percentile from the data. **(08 Marks)**

<i>Marks</i>	<i>No. of students</i>	<i>Marks</i>	<i>No. of students</i>
Less than 10	5	Less than 50	60
Less than 20	13	Less than 60	80
Less than 30	20	Less than 70	90
Less than 40	32	Less than 80	100

3. (a) Calculate the quartile deviation and its coefficient from the following data:

C.I.	0–10	10–20	20–30	30–40	40–50	50–60	60–70
<i>f</i>	18	25	33	42	38	71	23

(06 Marks)

- (b) Calculate the standard deviation from the following data: (06 Marks)

Income (Rs.)	10	20	30	40	50	60	70	80	90
No. of workers									

- (c) Calculate the Bowley's coefficient of skewness from the following data:

Wages	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89
No. of workers	5	9	14	20	25	15	8	4

(06 Marks)

4. (a) Define correlation. Explain the various types of correlation. (06 Marks)

- (b) Obtain the rank correlation coefficient for the following data: (06 Marks)

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

- (c) In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible: variance of $X = 9$. Regression equations $8X - 10Y = 0$, $40X - 18Y = 214$. Find on the basis of above information.

(i) The mean values of X and Y .(ii) The coefficient of correlation between X and Y .(iii) Standard deviation of Y . (08 Marks)

5. (a) Define the time series, with an example. List the various components of the time series. (06 Marks)

- (b) Calculate the trend values, by the method of moving average, assuming a four-yearly cycle, from the following data, related to sugar production in India:

Year	Sugar production (Lakh tonnes)	Year	Sugar Production (Lakh tonnes)
1971	37.4	1977	48.4
1972	31.1	1978	64.6
1973	38.7	1979	58.4
1974	39.5	1980	38.6
1975	47.9	1981	51.4
1976	42.6	1982	84.4

- (c) Calculate (i) Passche's index (ii) Laspeyre's index and (iii) Fisher's index numbers for the below given data: (08 Marks)

Commodity	Base Year		Current Year	
	Price	Quantity	Price	Quantity
M	10	20	12	25
N	8	13.5	10	22
O	20	8	25	10
P	18	8	20	7
Q	35	8	30	10

- 6.** (a) Explain the following terms, with suitable examples:

(i) Random experiment (ii) Sample space (iii) Mutually exclusive events.

(06 Marks)

- (b) In a bolt factory, machines A , B and C manufacture 25%, 35% and 40% respectively of the total. Of their output 5, 4 and 2 percent are defective bolts. A bolt is drawn at random from the production and is found to be defective. What is the probability that it was manufactured by machine B ?
- (c) Data was collected over a period of 10 years, showing number of deaths from horse kicks in each of the 200 army corps. The distribution of deaths was as follows:

No. of deaths	0	1	2	3	4	Total
Frequency	109	65	22	3	1	200

Fit a Poisson distribution to the data and calculate the theoretical frequencies.

(08 Marks)

- 7.** (a) Define sampling distribution. Write any four advantages of sampling.

(06 Marks)

- (b) The mean and variance of a random sample of 64 observations were computed as 160 and 100 respectively.
- (i) Compute the 95% confidence limits for population mean.
- (ii) If the investigator wants to be 95% confident that the error in estimate of population mean should not exceed ± 1.4 , how many additional observations are required.
- (c) Given the following information relating to two place A and B , test whether, there is any significant difference between their mean wages.

(06 Marks)

- (08 Marks)**

	<i>A</i>	<i>B</i>
Mean wages (Rs.)	47	49
Standard deviation (Rs.)	28	40
Number of workers	1000	1500

- 8.** (a) What is ANOVA? Explain the steps involved in carrying out Anova. **(06 Marks)**

- (b) The number of scooter accidents per month in a certain town were as follows:

12, 8, 20, 2, 14, 10, 15, 6, 9, 4

Are these frequencies in agreement with the belief that accident conditions were the same during this 10 month period? (Given $X^2_{0.05}$ for 9 d.f. = 16.919)

(06 Marks)

- (c) A certain stimulus administered to each of 12 patients resulted in the following changes in blood pressure:

5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4, 6

Can it be concluded that the stimulus will, in general, be accompanied by an increase in blood pressure?

(08 Marks)

MODEL QUESTION PAPER-IV
MBA Degree Examination
STATISTICS FOR MANAGEMENT

Time: 3 hrs.

Max. Marks: 100

- Note:** 1. Answer any **FIVE** full questions from Q. No. 1 to Q. No. 7.
 2. Question No. 8 is compulsory.

1. (a) Discuss briefly the different methods of sampling. Explain and illustrate their merits and demerits. **(07 Marks)**
- (b) In a certain examination the average grade of all students in class A is 68.4 and students in class B is 71.2. If the average of both classes combined is 70, find the ratio of the number of students in class A to the number of students in class B. **(05 Marks)**
- (c) The personnel manager of a factory wants to find a measure which he can use to fix the monthly income of the persons applying for a job in the production department. As an experimental project, he collected following data on 7 persons from that department, referring to years of service and their monthly income.

Years of service:	11	7	9	5	8	6	10
Income (in '000):	10	8	6	5	9	7	11

- (i) Fit the regression equation of income on years of service.
 (ii) Using it, what initial start would you recommend for a person applying for job, having served in similar capacity in another factory for 13 years? **(08 Marks)**

2. (a) What is meant by the theoretical frequency distribution? List out the properties of the binomial, Poisson and normal distributions. **(07 Marks)**
- (b) A company has two plants to manufacture scooters. Plant I manufactures 80% of the scooters and plant II manufactures 20%. At plant I, 85 out of 100 scooters are rated standard quality or better. At plant II, only 65 out of 100 scooters are rated standard quality or better. (i) What is the probability that a scooter selected at random came from plant I, if it is known that the scooter is of standard quality? (ii) What is the probability that a scooter selected at random came from plant II, if it is known that the scooter is of standard quality? **(05 Marks)**
- (c) A detergent maker has decided to change the appearance of the box. It has come up with four potential replacements, which are called, A, B, C and D. It shows the four designs to 400 randomly selected consumers, and ask them which design they like the best. Here are the results:

Design	A	B	C	D
Consumers	107	105	122	66

At the 0.05 level of significance, test the claim that all consumers like the four designs equally. **(08 Marks)**

3. (a) What are the desirable properties for an average? Under what circumstances would it be appropriate to use mean, mode, median and GM? Discuss. **(07 Marks)**
- (b) The following table gives indices of industrial production of registered unemployed (in hundred thousand). Calculate the value of the Karl Pearson's coefficient correlation.

Year	1991	1992	1993	1994	1995	1996	1997	1998
Index of production	100	102	104	107	105	112	103	99
No. unemployed	15	12	13	11	12	12	19	26

(05 Marks)

- (c) Draw an ogive for the following distribution. How many workers earned wages between Rs. 1365 and Rs. 1430? Also calculate the median wage.

Wages (Rs)	1000–1100	1100–1200	1200–1300	1400–1500	1500–1600	1600–1700
No. of workers	6	10	22	16	14	12

(08 Marks)

4. (a) What is an index number? Give Laspeyre's, Paasche's and Fisher's index numbers. Which one is the best and why? **(07 Marks)**

- (b) A market research firm is interested in surveying certain attitudes in a small community. There are 125 households broken down according to income, ownership of a telephone and ownership of a T.V.

	<i>Households with annual income of Rs. 8000 or less</i>		<i>Households with annual income above Rs. 8000</i>	
	<i>Telephone subscriber</i>	<i>No telephone subscriber</i>	<i>Telephone subscriber</i>	<i>No telephone subscriber</i>
Own TV set	27	20	18	10
No TV set	18	10	12	10

(08 Marks)

- (i) What is the probability of obtaining a TV owner in drawing at random?
(ii) What is the conditional probability of drawing a household that owns a TV, given that the household is a telephone subscriber? **(05 Marks)**

- (c) From the following data construct a price index number of the group of four commodities using the appropriate formula:

Commodity	<i>Base Year</i>		<i>Current Year</i>	
	<i>Price for unit (Rs.)</i>	<i>Expenditure (Rs.)</i>	<i>Price for unit (Rs.)</i>	<i>Expenditure (Rs.)</i>
A	2	40	5	75
B	4	16	8	40
C	1	10	2	24
D	5	25	10	60

(08 Marks)

5. (a) Briefly explain the various methods of determining trend in a time series. Explain the merits and demerits of each method. **(07 Marks)**

- (b) The coefficient of rank correlation of the marks obtained by 10 students in statistics and accountancy was found to be 0.2. It was later discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 9 instead of 7. Find the correct coefficient of rank correlation. **(05 Marks)**

- (c) Below are given the annual production (in thousand tonnes) of a fertilizer factory:

Years	1997	1998	1999	2000	2001	2002	2003
Production	70	75	90	91	95	98	100

Fit a straight line trend by the method of least squares and tabulate the trend values. **(08 Marks)**

6. (a) Show clearly the necessity and importance of diagrams in statistics. What precautions should be taken in drawing a good diagram? **(07 Marks)**
- (b) The heights of adult females are normally distributed with a mean of 63.6 inches and a standard deviation of 2.5 inches. Find the probability that a randomly selected female is 5 feet tall or shorter. **(05 Marks)**
- (c) The following data represent the number of units of production per day, tuned out by 5 different workers, using 4 different types of machines:

Workers	Machine type			
	A	B	C	D
1	44	38	47	36
2	46	40	52	43
3	34	36	44	32
4	43	38	46	33
5	38	42	49	39

Test whether the mean productivity is the same for different machines. Test whether the 5 men differ with respect to mean productivity. **(08 Marks)**

7. (a) What is coefficient of variation? What purpose does it serve? Also distinguish between 'variance' and 'coefficient of variation'. **(07 Marks)**
- (b) The number of employees, wages per employee and the variance of the wages per employee for two factories are given below:

	Factory A	Factory B
No. of employees	100	150
Average wage per employee per month (Rs.)	3200	2800
Variance of the wages per employee per month (Rs.)	625	729

In which factory is there greater variation in the distribution of wages per employee? **(05 Marks)**

- (c) A statistics professor has designed a final exam that he believes will produce a mean score of 70. Mr. Sriram, one of his colleagues, disagrees, claiming that the mean score for all statistics students in this exam will be below 70. Mr. Sriram randomly selects 38 statistics students and gives them the exam. Here are their scores:

32	41	43	44	46	46	47	48	49	52
54	54	56	57	58	58	60	60	65	66
66	67	67	68	68	68	68	72	76	83
87	92	92	94	100	50	61	71		

Use this sample data to test Mr. Sriram's claim at the 0.01 level of significance. **(08 Marks)**

8. (a) Explain the procedure generally followed in testing of a hypothesis. Point out the difference between one tailed and two tailed tests. **(07 Marks)**
- (b) Calculate geometric mean from the following data: **(05 Marks)**

125	1462	38	7	0.22	0.08	12.75	0.5	
-----	------	----	---	------	------	-------	-----	--

- (c) 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 defectives	0	1	2	3	4	5	6	7
No. of samples	7	6	19	35	30	23	7	1

Fit a binomial distribution and find the expected frequencies if the chance of machine being defective is $1/2$.

This page is intentionally left blank.

SOLUTION TO MODEL QUESTION PAPER-I

1. (a) Classification of data is the process of arranging data in groups/classes on the basis of certain properties. The classification of statistical data serves the following purposes:
 - (i) It condenses the raw data into a form suitable for statistical analysis.
 - (ii) It removes complexities and highlights the features of the data.
 - (iii) It facilitates comparisons and drawing inferences from the data. For example, if university students in a particular course are divided according to sex, their results can be compared.
 - (iv) It provides information about the mutual relationships among elements of a data set. For example, based on literacy and criminal tendency of a group of people, it can be established whether literacy has any impact or not on criminal tendency.
 - (v) It helps in statistical analysis by separating elements of the data set into homogeneous groups and hence brings out the points of similarity and dissimilarity.

Requisites of Ideal Classification: The classification of data is decided after taking into consideration the nature, scope, and purpose of the investigation. However, an ideal classification should have following characteristics:

It should be unambiguous It is necessary that the various classes should be so defined that there is no room for confusion. There must be only one class for each element of the data set. For example, if the population of the country is divided into two classes, say literates and illiterates, then an exhaustive definition of the terms used would be essential.

Classes should be exhaustive and mutually exclusive Each element of the data set must belong to a class. For this, an extra class can be created with the title 'others' so as to accommodate all the remaining elements of the data set.

Each class should be mutually exclusive so that each element must belong to only one class.

It should be stable The classification of a data set into various classes must be done in such a manner that if each time an investigation is conducted, it remains unchanged and hence the results of one investigation may be compared with that of another. For example, classification of the country's population by a census survey based on occupation suffers from this defect because various occupations are defined in different ways in successive censuses and, as such, these figures are not strictly comparable.

It should be flexible A classification should be flexible so that suitable adjustments can be made in new situations and circumstances. However, flexibility does not mean instability. The data should be divided into few major classes which must be further subdivided. Ordinarily there would not be many changes in the major classes. Only small sub-classes may need a change and the classification can thus retain the merit of stability and yet have flexibility.

The term stability does not mean rigidity of classes. The term is used in a relative sense. One-time classification can not remain stable forever. With change in time, some classes become obsolete and have to be dropped and fresh classes have to be added. The classification may be called ideal if it can adjust itself to these changes and yet retain its stability.

Basis of Classification Statistical data are classified after taking into account the nature, scope, and purpose of an investigation. Generally, data are classified on the basis of the following four bases:

Geographical Classification In geographical classification, data are classified on the basis of geographical or locational differences such as—cities, districts, or villages between various elements of the data set.

Such a classification is also known as *spatial classification*. Geographical classifications are generally listed in alphabetical order. Elements in the data set are also listed by the frequency size to emphasize the importance of various geographical regions as in ranking the metropolitan cities by population density. The first approach is followed in case of reference tables while the second approach is followed in the case of summary tables.

Chronological Classification When data are classified on the basis of time, the classification is known as chronological classification. Such classifications are also called *time series* because data are usually listed in chronological order starting with the earliest period.

Qualitative Classification In qualitative classification, data are classified on the basis of descriptive characteristics or on the basis of attributes like sex, literacy, region, caste, or education, which cannot be quantified. This is done in two ways:

- (i) *Simple classification*: In this type of classification, each class is subdivided into two sub-classes and only one attribute is studied such as: male and female; blind and not blind, educated and uneducated, and so on.
- (ii) *Manifold classification*: In this type of classification, a class is subdivided into more than two sub-classes which may be sub-divided further. An example of this form of classification is shown in the box:

Quantitative Classification In this classification, data are classified on the basis of some characteristics which can be measured such as height, weight, income, expenditure, production, or sales.

Quantitative variables can be divided into the following two types. The term variable refers to any quantity or attribute whose value varies from one investigation to another.

- (i) *Continuous variable* is the one that can take any value within the range of numbers. Thus the height or weight of individuals can be of any value within the limits. In such a case, data are obtained by measurement.
- (ii) *Discrete* (also called *discontinuous*) *variable* is the one whose values change by steps or jumps and can not assume a fractional value. The number of children in a family, number of workers (or employees), number of students in a class, are few examples of a discrete variable. In such a case data are obtained by counting.

1. (b) Let \bar{x}_1, \bar{x}_2 be the average daily wages of male and female workers \bar{x} be the average daily wages of all the workers in a factory respectively.

Given: $\bar{x} = 444$

$$\bar{x}_1 = 480$$

$$\bar{x}_2 = 360$$

Total wages of male workers = $480 n_1$

Total wages of female workers = $360 n_2$

Total wages of all workers = $444 (n_1 + n_2)$

$$\therefore n_1 \bar{x}_1 + n_2 \bar{x}_2 = (n_1 + n_2) \bar{x}$$

$$480 n_1 + 360 n_2 = 444 (n_1 + n_2)$$

$$480 n_1 + 360 n_2 = 444 n_1 + 444 n_2$$

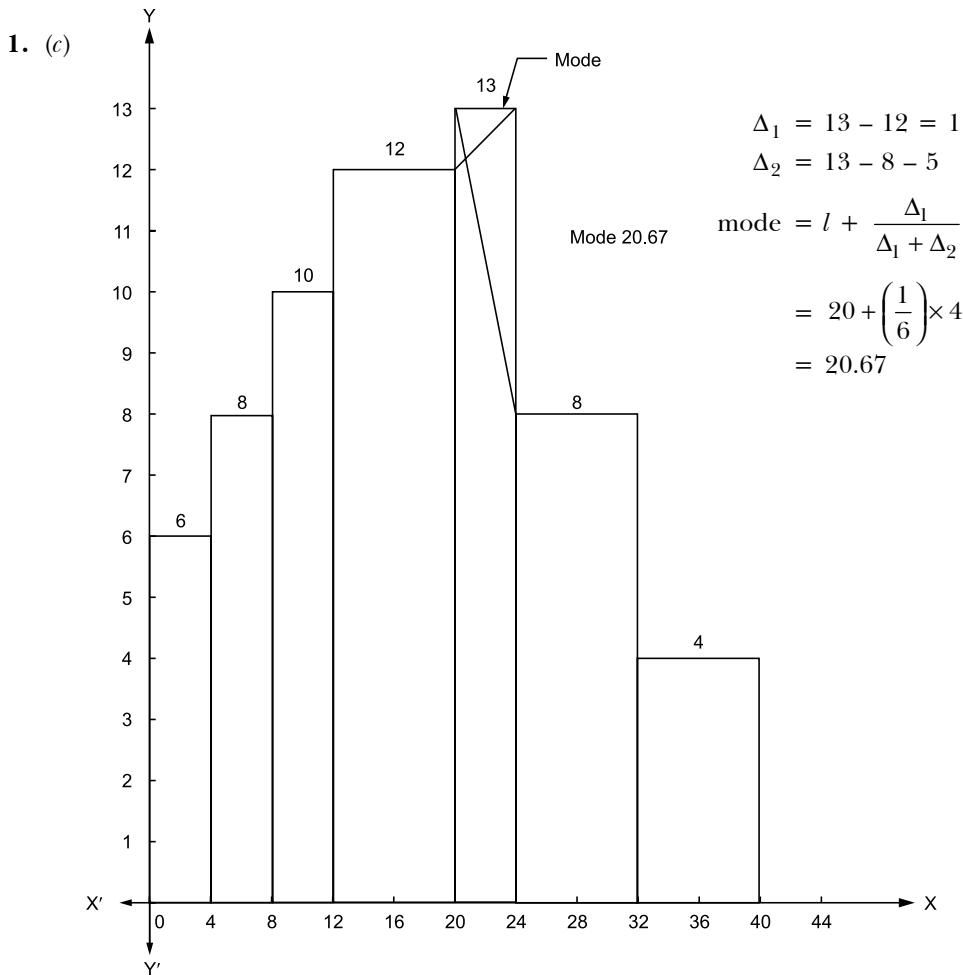
$$480 n_1 - 444 n_1 = 444 n_2 - 360 n_2$$

$$36 n_1 = 84 n_2 \Rightarrow \frac{n_1}{n_2} = \frac{84}{36}$$

$$\Rightarrow n_1 = 84 \text{ and } n_2 = 36$$

$$\therefore \text{Percentage of male employees} = \frac{n_1}{n_1 + n_2} \times 100 = \frac{84}{120} \times 100 = 70\%$$

$$\text{Percentage of female employees} = \frac{n_2}{n_1 + n_2} \times 100 = \frac{36}{120} \times 100 = 30\%$$



2. (a) (i) Individuals, focus groups, and/or panels of respondents specifically decided upon and set up by the investigator for data collection are examples of primary data sources. Any one or a combination of the following methods can be chosen to collect primary data:

- (i) Direct personal observations
- (ii) Direct or indirect oral interviews
- (iii) Administering questionnaires

Government publications, which include

- (i) The National Accounts Statistics, published by the Central Statistical Organization (CSO). It contains estimates of national income for several years, growth rate, and rate on major economic activities such as agriculture, industry, trade, transport, and so on;
- (ii) Wholesale Price Index, published by the office of the Economic Advisor, Ministry of Commerce and Industry;
- (iii) Consumer Price Index;
- (iv) Reserve Bank of India bulletins;
- (v) Economic Survey.

Non-Government publications include publications of various industrial and trade associations such as

- (i) The Indian Cotton Mills Association

2. (a) (ii) The process of selecting a sample from a population is called *sampling*. In sampling, a representative *sample* or *portion* of elements of a population or process is selected and then analysed. Based on sample results, called *sample statistics*,

statistical inferences are made about the population characteristic. For instance, a political analyst selects specific or random set of people for interviews to estimate the proportion of the votes that each candidate may get from the population of voters; an auditor selects a sample of vouchers and calculates the sample mean for estimating population average amount; or a doctor examines a few drops of blood to draw conclusions about the nature of disease or blood constitution of the whole body.

2. (b) Spearman's Rank Correlation Co-efficient:

x	y	R_1	R_2	$d = R_1 - R_2$	d^2
39	47	8	10	-2	4
65	53	6	8	-2	4
62	58	7	7	0	0
90	86	2	2	0	0
82	62	3	5	-2	4
75	68	5	4	1	1
25	60	10	6	4	16
98	91	1	1	0	0
36	51	9	9	0	0
78	84	4	3	1	1
					<u>30</u>

$$\begin{aligned} P &= 1 - \left[\frac{6\sum d^2}{N(N^2-1)} \right] = 1 - \left(\frac{6(30)}{10 \times 99} \right) \\ &= 1 - \left[\frac{180}{990} \right] = 1 - 0.1818 \end{aligned}$$

$$P = 0.82$$

2. (c) Karl Pearson's Co-efficient of Skewness:

No. of accidents x	No. of Cities f	$(x-12)$ d	fd	fd^2
10	2	-2	-4	8
11	4	-1	-4	4
12	10	0	0	0
13	8	1	8	8
14	5	2	10	20
15	1	3	3	9
$N = 30$			<u>13</u>	<u>49</u>

$$S_{kp} = \frac{\text{Mean} - \text{Mode}}{\text{SD}}$$

$$\text{Mean} = A + \frac{\Sigma fd}{N} = 12 + \left(\frac{13}{30} \right) = 12 + 0.43 = 12.43$$

Mode = 12 (Max. frequency)

$$\begin{aligned} \text{SD} &= \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N} \right)^2} = \sqrt{\left(\frac{49}{30} \right) - \left(\frac{13}{30} \right)^2} \\ &= \sqrt{(1.63) - (0.43)^2} = \sqrt{1.63 - 0.1849} = \sqrt{1.4451} = 1.20 \end{aligned}$$

$$S_{kp} = \frac{\text{Mean} - \text{Mode}}{\text{SD}} = \frac{12.43 - 12}{1.20} = \frac{0.43}{1.20} = 0.36$$

3. (a) The limitations or disadvantages of the range can partially be overcome by using another measure of variation which measures the spread over the middle half of the values in the data set so as to minimise the influence of outliers (extreme values) in the calculation of range. Since a large number of values in the data set lie in the central part of the frequency distribution, therefore it is necessary to study the **Interquartile Range** (also called midspread). To compute this value, the entire data set is divided into four parts each of which contains 25 per cent of the observed values. The quartiles are the highest values in each of these four parts. The *interquartile range* is a measure of dispersion or spread of values in the data set between the third quartile, Q_3 and the first quartile, Q_1 . In other words, the *interquartile range or deviation* (IQR) is the range for the middle 50 per cent of the data. The concept of IQR is shown below:

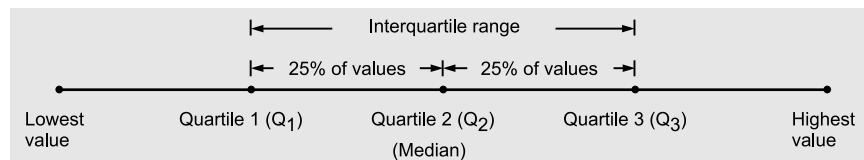
$$\text{Interquartile range (IQR)} = Q_3 - Q_1$$

Half the distance between Q_1 and Q_3 is called the *semi-interquartile range* or the *quartile deviation* (QD).

$$\text{Quartile deviation (QD)} = \frac{Q_3 - Q_1}{2}$$

The median is not necessarily midway between Q_1 and Q_3 , although this will be so for a symmetrical distribution. The median and quartiles divide the data into equal numbers of values but do not necessarily divide the data into equally wide intervals.

As shown above the quartile deviation measures the average range of 25 per cent of the values in the data set. It represents the spread of all observed values because its value is computed by taking an average of the middle 50 per cent of the observed values rather than of the 25 per cent part of the values in the data set.



In a non-symmetrical distribution, the two quartiles Q_1 and Q_3 are at equal distance from the median, that is, $\text{Median} - Q_1 = Q_3 - \text{Median}$. Thus, $\text{Median} \pm \text{Quartile Deviation}$ covers exactly 50 per cent of the observed values in the data set.

A smaller value of quartile deviation indicates high uniformity or less variation among the middle 50 per cent observed values around the median value. On the other hand, a high value of quartile deviation indicates large variation among the middle 50 per cent observed values.

Coefficient of Quartile Deviation: Since quartile deviation is an absolute measure of variation, therefore its value gets affected by the size and number of observed values in the data set. Thus, the Q.D. of two or more than two sets of data may differ. Due to this reason, to compare the degree of variation in different sets of data, we compute the relative measure corresponding to Q.D., called the *coefficient of Q.D.*, and it is calculated as follows:

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Advantages and Disadvantages of Quartile Deviation The major advantages and disadvantages of quartile deviation are summarized as follows:

Advantages

- (i) It is not difficult to calculate but can only be used to evaluate variation among observed values within the middle of the data set. Its value is not affected by the extreme (highest and lowest) values in the data set.
- (ii) It is an appropriate measure of variation for a data set summarized in open-ended class intervals.

- (iii) Since it is a positional measure of variation, therefore it is useful in case of erratic or highly skewed distributions, where other measures of variation get affected by extreme values in the data set.

Disadvantages

- (i) The value of Q.D. is based on the middle 50 per cent observed values in the data set, therefore it cannot be considered as a good measure of variation as it is not based on all the observations.
- (ii) The value of Q.D. is very much affected by sampling fluctuations.
- (iii) The Q.D. has no relationship with any particular value or an average in the data set for measuring the variation. Its value is not affected by the distribution of the individual values within the interval of the middle 50 per cent observed values.

3. (b) Calculate Laspeyre's, Paasche's and Fisher's index numbers

Item	p_0	p_1	q_0	q_1	$q_1 p_0$	$q_0 p_0$	$q_1 p_1$	$q_0 p_1$
A	6	10	300	560	3360	1800	5600	3000
B	2	2	200	240	480	400	480	400
C	4	6	240	360	1440	960	2160	1440
D	10	12	300	288	2880	3000	3456	3600
E	8	12	320	432	3456	2560	5184	3840
					11616	8720	16880	12280

Laspeyre's Index number:

$$p_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{12280}{8720} \times 100 = 140.83$$

Paasche's Index:

$$p_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{16880}{11616} \times 100 = 145.32$$

Fisher's Index:

$$\begin{aligned} p_{01} &= \sqrt{L \times P} && L \rightarrow \text{Laspeyre's index} \\ &= \sqrt{(140.83)(145.32)} && P \rightarrow \text{Paasche's index} \\ &= \sqrt{20465.42} = 143.06 \end{aligned}$$

3. (c) For the given data find the missing values

$$\text{Mode} = l + \left[\frac{f_1 - f_0}{2f_1 - (f_0 + f_2)} \right] \times h = 45 + \left(\frac{18 - x}{36 - (x + 14)} \right) \times 5$$

$$\Rightarrow \text{Given Mode} = 47.5$$

$$\Rightarrow \frac{(18 - x) 5}{36 - 14 - x} = 47.5 - 45$$

$$\frac{90 - 5x}{22 - x} = 2.5$$

$$90 - 5x = 2.5 (22 - x)$$

$$\Rightarrow 2.5x = 35$$

$$x = \frac{35}{2.5}$$

$$x = 14$$

$$x + y = 18 \Rightarrow y = 4$$

\therefore Missing values are:

$$x = 14$$

$$y = 4$$

4. (a) (i) A single possible outcome (or result) of an experiment is called a simple (or elementary) event. An **event** is the set (or collection) of one or more simple events of an experiment in the sample space and having a specific common characteristic.
- (ii) **Random experiment:** A process of obtaining information through observation or measurement of a phenomenon whose outcome is subject to chance.
- (iii) **Sample space:** The set of all possible outcomes or simple events of an experiment.
- (iv) Two or more events are said to be equally likely if each has an equal chance to occur. That is, one of them cannot be expected to occur in preference to the other. For example, each number may be expected to occur on the uppermost face of a rolling die the same number of times in the long run.
- (v) A general definition of probability states that **probability** is a numerical measure (between 0 and 1 inclusively) of the likelihood or chance of occurrence of an uncertain event. However, it does not tell us how to compute the probability.
- (vi) The addition rules are helpful when we have two events and are interested in knowing the probability that at least one of the events occurs.

4. (b) Event A – getting plumbing contract
 Event B – getting electrical contract

$$\begin{aligned}
 P(A) &= \frac{2}{3} \\
 P(B^c) &= \frac{5}{9} \Rightarrow P(B) = \frac{4}{9} \\
 P(A \cup B) &= \frac{4}{5} \\
 P(A \cup B) &= ? \\
 P(A \cap B) &= P(A) + P(B) - P(A \cap B) \\
 \Rightarrow P(A \cap B) &= \frac{2}{3} + \frac{4}{9} - \frac{4}{5} \\
 &= 0.67 + 0.44 - 0.8 = 0.31
 \end{aligned}$$

The probability that the contractor will get both contract is **0.31**.

4. (c) **Linear trend:**

Year t	Exports y	$x = t - 1992$	x^2	xy
1990	11	-2	4	-22
1991	16	-1	1	-16
1992	13	0	0	0
1993	18	1	1	18
1994	22	2	4	44
1995	20	3	9	60
	$\Sigma y = 100$	$\Sigma x = 3$	$\Sigma x^2 = 19$	$\Sigma xy = 84$

Normal equations:

$$\begin{aligned}
 (i) \Sigma y &= na + b\Sigma x \\
 100 &= 6a + 3b \quad \dots(1)
 \end{aligned}$$

$$\begin{aligned}
 (ii) \Sigma xy &= a\Sigma x + b\Sigma x^2 \\
 84 &= 3a + 19b \quad \dots(2)
 \end{aligned}$$

Solving (1) and (2)

$$(1) \rightarrow 6a + 3b = 100$$

$$(2) \rightarrow 3a + 19b = 84$$

$$\begin{array}{rcl}
 (1) \rightarrow & 6a + 3b = 100 \\
 (2) \times 2 \rightarrow & 6a + 38b = 168 \\
 & (-) \quad (-) \quad (-) \\
 & \hline
 & -35b = -68 \\
 b & = \frac{68}{35} \\
 b & = 1.94
 \end{array}$$

Substitute b in (1)

$$\begin{aligned}
 \Rightarrow \quad 6a + 3(1.94) &= 100 \\
 6a + 5.82 &= 100 \\
 6a &= 100 - 5.82 \\
 6a &= 94.18 \\
 a &= \frac{94.18}{6} \\
 a &= 15.70
 \end{aligned}$$

\therefore Linear trend fitted to yearly values is:

$$Y = a + bx$$

$$\text{i.e.} \quad Y = 15.70 + 1.94X$$

When $X = 13$

$$Y_{2005} = 15.70 + 1.94(13)$$

$$Y = 15.70 + 25.22$$

$$\mathbf{Y = 40.92 \text{ crores}}$$

5. (a) A statistical technique that is used to analyse the strength and direction of the relationship between two quantitative variables, is called *correlation analysis*.

There are three broad types of correlations:

1. Positive and negative,
2. Linear and non-linear,
3. Simple, partial, and multiple.

In this chapter we will discuss simple linear positive or negative correlation analysis.

Positive and Negative Correlation A positive (or direct) correlation refers to the same direction of change in the values of variables. In other words, if values of variables are varying (i.e., increasing or decreasing) in the same direction, then such correlation is referred to as **positive correlation**.

A **negative (or inverse) correlation** refers to the change in the values of variables in opposite direction.

Linear and Non-Linear Correlation A linear correlation implies a constant change in one of the variable values with respect to a change in the corresponding values of another variable. In other words, a correlation is referred to as *linear correlation* when variations in the values of two variables have a constant ratio.

When these pairs of values of x and y are plotted on a graph paper, the line joining these points would be a straight line.

A non-linear (or curvi-linear) correlation implies an absolute change in one of the variable values with respect to changes in values of another variable. In other words, a correlation is referred to as a *non-linear correlation* when the amount of change in the values of one variable does not bear a constant ratio to the amount of change in the corresponding values of another variable.

When these pair of values of x and y are plotted on a graph paper, the line joining these points would not be a straight line, rather it would be curvi-linear.

Simple, Partial, and Multiple Correlation The distinction between simple, partial, and multiple correlation is based upon the number of variables involved in the correlation analysis.

If only two variables are chosen to study correlation between them, then such a correlation is referred to as *simple correlation*. A study on the yield of a crop with respect to only amount of fertilizer, or sales revenue with respect to amount of money spent on advertisement, are a few examples of simple correlation.

In *partial correlation*, two variables are chosen to study the correlation between them, but the effect of other influencing variables is kept constant. For example (i) yield of a crop is influenced by the amount of fertilizer applied, rainfall, quality of seed, type of soil, and pesticides, (ii) sales revenue from a product is influenced by the level of advertising expenditure, quality of the product, price, competitors, distribution, and so on. In such cases an attempt to measure the correlation between yield and seed quality, assuming that the average values of other factors exist, becomes a problem of partial correlation.

In *multiple correlation*, the relationship between more than three variables is considered simultaneously for study. For example, employer-employee relationship in any organization may be examined with reference to, training and development facilities; medical, housing, and education to children facilities; salary structure; grievances handling system; and so on.

5. (b) Compute the median and 63rd percentile

Weight (in kgs)	No. of bags <i>f</i>	<i>cf</i>
0–4.5	5	5
4.5–9.5	7	12
9.5–14.5	10	22
14.5–19.5	8	30
19.5–24.5	6	36
24.5–29.5	4	40
	40	

$$\begin{aligned}
 \text{Median} &= l + \left\{ \frac{h}{f} \left(\frac{N}{2} - C \right) \right\} & N &= 40 \\
 &= 9.5 + \left\{ \frac{5}{10} (20 - 12) \right\} & \frac{N}{2} &= 20 \\
 &= 9.5 + \{0.5(8)\} & l &= 9.5 \\
 &= 9.5 + 4 = 13.5 & h &= 5 \\
 & & f &= 10 \\
 & & C &= 12
 \end{aligned}$$

63rd percentile:

$$\text{Percentile } 63^{\text{rd}} \text{ item} = \frac{63N}{100} = 25.2$$

The cumulative frequency just greater than 25.2 is 30.

Hence the corresponding class 14.5 – 19.5 contains P_{63}

$$\begin{aligned}
 \therefore P_{63} &= 14.5 + \left[\frac{25.2 - 22}{8} \right] \times 5 \\
 &= 14.5 + (0.4) \times 5 \\
 &= 14.5 + 2 \\
 \mathbf{P}_{63} &= \mathbf{16.5}
 \end{aligned}$$

5. (c) Regression Lines:

X	Y	X^2	Y^2	XY
65	67	4225	4489	4355
66	68	4356	4624	4488
67	64	4489	4096	4288
67	68	4489	4624	4556
68	72	4624	5184	4896
69	70	4761	4900	4830
71	69	5041	4761	4899
73	70	5329	4900	5110
546	548	37314	37578	37422

$$\bar{X} = \frac{\Sigma x}{n} = \frac{546}{8} = 68.25$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{548}{8} = 68.5$$

Equation for Y on X is: $Y - \bar{Y} = b_{YX}(X - \bar{X})$

$$\begin{aligned} b_{YX} &= \frac{N \Sigma XY - \Sigma X \Sigma Y}{N \Sigma X^2 - (\Sigma X)^2} = \frac{8(37422) - (546)(548)}{8(37314) - (546)^2} \\ &= \frac{299376 - 299208}{298512 - 298116} = \frac{168}{396} \\ b_{YX} &= 0.42 \end{aligned}$$

Equation for X on Y is: $X - \bar{X} = b_{XY}(Y - \bar{Y})$

$$\begin{aligned} b_{XY} &= \frac{N \Sigma XY - \Sigma X \Sigma Y}{N \Sigma Y^2 - (\Sigma Y)^2} = \frac{8(37422) - (546)(548)}{8(37578) - (548)^2} \\ &= \frac{299376 - 299208}{300624 - 300304} = \frac{168}{320} \\ b_{XY} &= 0.52 \end{aligned}$$

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

$$Y - 68.5 = 0.42(X - 68.25)$$

$$Y - 68.5 = 0.42 X - 28.66$$

$$Y = 0.42 X - 28.66 + 68.5$$

$$Y = 0.42 X + 39.84$$

The required equation is:

$$Y = 0.42X + 39.84$$

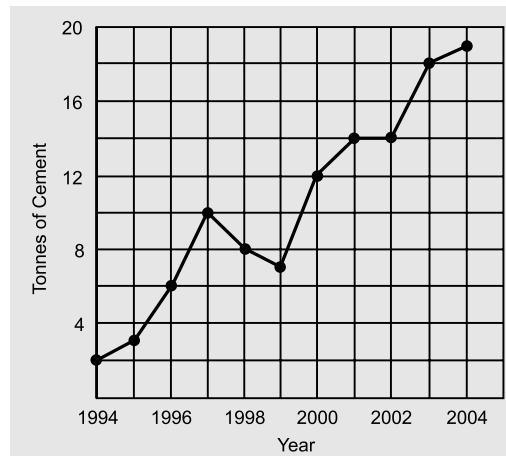
When $X = 67.5$,

$$Y = 0.42(67.5) + 39.84 = 28.35 + 39.84$$

$$Y = 68.19$$

6. (a) A time series is a set of numerical values of some variable obtained at regular period over time. The series is usually tabulated or graphed in a manner that readily conveys the behaviour of the variable under study. The below figure presents the export of cement (in tonnes) by a cement company between 1994 and 2004. The graph suggests that the series is time dependent. The management of the company is interested in determining how the series is dependent on time and in developing a means of predicting future levels with some degree of reliability. The nature of the time dependence is often analysed by decomposing the time series into its components.

Year	Export (tonnes)
1994	2
1995	3
1996	6
1997	10
1998	8
1999	7
2000	12
2001	14
2002	14
2003	18
2004	19



6. (b) Mean deviation and its co-efficient using Median

C.I.	Mid point X	F	cf	fx	(X-M)	f X-M
0–10	5	8	8	40	36.7	293.6
10–20	15	15	23	225	26.7	400.5
20–30	25	24	47	600	16.7	400.8
30–40	35	34	81	1190	6.7	227.8
40–50	45	47	128	2115	3.3	155.1
50–60	55	50	178	2750	13.3	665
		178		6920		2142.8

$$\text{Median} = l + \left\{ \frac{h}{f} \left(\frac{N}{2} - C \right) \right\} = 40 + \left\{ \frac{10}{47} (89 - 81) \right\} = 40 + (0.2128(8)) \\ = 40 + (1.7024) = 41.7024 \approx 41.7 \quad l = 40$$

$$MD = \frac{\sum f|x-M|}{N} = \frac{2142.8}{178} = 12.038 \approx 12.04 \quad h = 10 \\ f = 47$$

$$\text{Co-efficient of Mean deviation about Mdian} = \frac{MD}{\text{Median}} = \frac{12.04}{41.7} \\ = 0.2887 \approx 0.29 \quad C = 81$$

6. (c) Poisson distribution:

x	f	fx
0	123	0
1	59	59
2	14	28
3	3	9
4	1	4
	200	100

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{100}{200}$$

$$\bar{X} = 0.5 = \lambda, \quad N = 200 = \sum f$$

$$\frac{e^{-\lambda} \lambda^x}{x!}$$

X	$E = N.P(X = x)$	E
0	121.3 ~ 121	121
1	60.65 ~ 61	61
2	15.16 ~ 15	15
3	2.5 ~ 3	3
4	0.3 ~ 0.3	0
		200

X	0	1	2	3	4
Expected Frequencies E	121	61	15	3	0

7. (a) Sampling methods compared to census provides an attractive means of learning about a population or process in terms of reduced cost, time and greater accuracy. The representation basis and the element selection techniques from the given population, classify several sampling methods into two categories.

Types of Sampling Methods

<i>Element Selection</i>	<i>Representation Basis</i>	
	<i>Probability (Random)</i>	<i>Non-probability (Non-random)</i>
• Unrestricted	Simple random sampling	Convenience sampling
• Restricted	Complex random sampling • Stratified sampling • Cluster sampling • Systematic sampling • Multi-stage sampling	Purposive sampling • Quota sampling • Judgement sampling

Probability Sampling Methods Several probability sampling methods for selecting samples from a population or process are as follows:

Simple Random (Unrestricted) Sampling In this method, every member (or element) of the population has an equal and independent chance of being selected again and again when a sample is drawn from the population. To draw a random sample, we need a complete list of all elements in the population of interest so that each element can be identified by a distinct number. Such a list is called *frame for experiment*. The frame for experiment allows us to draw elements from the population by randomly generating the numbers of the elements to be included in the sample.

This method is suitable for sampling, as many statistical tests assume independence of sample elements. One disadvantage with this method is that all elements of the population have to be available for selection, which many a times is not possible.

Stratified Sampling This method is useful when the population consists of a number of heterogeneous subpopulations and the elements within a given subpopulation are relatively homogeneous compared to the population as a whole. Thus, population is divided into mutually exclusive groups called *strata* that are relevant, appropriate and meaningful in the context of the study. A simple random sample, called a *sub-sample*, is then drawn from each *strata* or *group*, in proportion or a non-proportion to its size. As the name implies, a proportional sampling procedure requires that the number of elements in each stratum be in the same proportion as in the population. In non-proportional procedure, the number of elements in each stratum are disproportionate to the respective numbers in the population.

The basis for forming the strata such as location, age, industry type, gross sales, or number of employees, is at the discretion of the investigator. Individual stratum samples are combined into one to obtain an overall sample for analysis.

This sampling procedure is more efficient than the simple random sampling procedure because, for the same sample size, we get more representativeness from each important segment of the population and obtain more valuable and differentiated information with respect to each strata.

Cluster Sampling This method, sometimes known as *area sampling method*, has been devised to meet the problem of costs or inadequate sampling frames (a complete listing of all elements in the population so that each member can be identified by a distinct number). The entire population to be analysed is divided into smaller groups or chunks of elements and a sample of the desired number of areas selected by a simple random sampling method. Such groups are termed as *clusters*. The elements of a cluster are called *elementary units*. These clusters do not have much heterogeneity among the elements. A household where individuals live together is an example of a cluster.

If several groups with intragroup heterogeneity and intergroup homogeneity are found, then a random sampling of the clusters or groups can be done with information gathered from each of the elements in the randomly chosen clusters. Cluster samples offer more heterogeneity within groups and more homogeneity among groups—the reverse of what we find in stratified random sampling, where there is homogeneity within each group and heterogeneity across groups.

For instance, committees formed from various departments in an organization to offer inputs to make decisions on product development, budget allocations, marketing strategies, etc are examples of different clusters. Each of these clusters or groups contains a heterogeneous collection of members with different interests, orientations, values, philosophy, and vested interests. Based on individual and combined perceptions, it is possible to make final decision on strategic moves for the organization.

In summary, cluster sampling involves preparing only a list of clusters instead of a list of individual elements. For examples, (i) residential blocks (colonies) are commonly used to cluster in surveys that require door-to-door interviews, (ii) airlines sometimes select randomly a set of flights to distribute questionnaire to every passenger on those flights to measure customer satisfaction. In this situation, each flight is a cluster. It is much easier for the airline to choose a random sample of flights than to identify and locate a random sample of individual passengers to distribute questionnaire.

Multistage Sampling This method of sampling is useful when the population is very widely spread and random sampling is not possible. The researcher might stratify the population in different regions of the country, then stratify by urban and rural and then choose a random sample of communities within these strata. These communities are then divided into city areas as clusters and randomly consider some of these for study. Each element in the selected cluster may be contacted for desired information.

For example, for the purpose of a national pre-election opinion poll, the *first stage* would be to choose as a sample a specific state (region). The size of the sample, that is the number of interviews, from each region would be determined by the relative populations in each region. In the *second stage*, a limited number of towns/cities in each of the regions would be selected, and then in the *third stage*, within the selected towns/cities, a sample of respondents could be drawn from the electoral roll of the town/city selected at the second stage.

The essence of this type of sampling is that a subsample is taken from successive groups or strata. The selection of the sampling units at each stage may be achieved with or without stratification. For example, at the second stage when the sample of towns/cities is being drawn, it is customary to classify all the urban areas in

the region in such a way that the elements (towns/cities) of the population in those areas are given equal chances of inclusion.

Systematic Sampling This procedure is useful when elements of the population are already physically arranged in some order, such as an alphabetized list of people with driving licenses, list of bank customers by account numbers. In these cases one element is chosen at random from first k element and then every k th element (member) is included in the sample. The value k is called the *sampling interval*. For example, suppose a sample size of 50 is desired from a population consisting of 100 accounts receivable. The sampling interval is $k = N/n = 1000/50 = 20$. Thus a sample of 50 accounts is identified by moving systematically through the population and identifying every 20th account after the first randomly selected account number.

Non-Random Sampling Methods Several non-random sampling methods for selecting samples from a population or process are as follows:

Convenience Sampling In this procedure, units to be included in the sample are selected at the convenience of the investigator rather than by any prespecified or known probabilities of being selected. For example, a student for his project on 'food habits among adults' may use his own friends in the college to constitute a sample simply because they are readily available and will participate for little or no cost. Other examples are, public opinion surveys conducted by any TV channel near the railway station; bus stop, or in a market.

Convenience samples are easy for collecting data on a particular issue. However, it is not possible to evaluate its representativeness of the population and hence precautions should be taken in interpreting the results of convenient samples that are used to make inferences about a population.

Purposive Sampling Instead of obtaining information from those who are most conveniently available, it sometimes becomes necessary to obtain information from specific targets—respondents who will be able to provide the desired information either because they are the only ones who can give the desired information or because they satisfy to some criteria set by researcher.

Judgement Sampling Judgement sampling involves the selection of respondents who are in the best position to provide the desired information. The judgment sampling is used when a limited number of respondents have the information that is needed. In such cases, any type of probability sampling across a cross section of respondents is purposeless and not useful. This sampling method may curtail the generalizability of the findings due to the fact that we are using a sample of respondents who are conveniently available to us. It is the only viable sampling method for obtaining the type of information that is required from very specific section of respondents who possess the knowledge and can give the desired information.

However, the validity of the sample results depend on the proper judgment of the investigator in choosing the sample. Great precaution is needed in drawing conclusions based on judgment samples to make inferences about a population.

Quota Sampling Quota Sampling is a form of proportionate stratified sampling in which a predetermined proportion of elements are sampled from different groups in the population, but on convenience basis. In other words, in quota sampling the selection of respondents lies with the investigator, although in making such selection he/she must ensure that each respondent satisfies certain criteria which is essential for the study.

7. (b) (i) Null hypothesis $H_0: \mu_1 + \mu_2$ i.e. Meditation does not reduce BP

Alternative hypothesis: $H_1: \mu_1 < \mu_2$ i.e. Meditation reduces BP

(ii) Computation of Test Statistic:

Under H_0 , the test statistic is:

$$t = \frac{\bar{d}}{\sqrt{\frac{S^2}{n-1}}} \sim t_{n-1}$$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{-38}{8} = -4.75$$

$$\begin{aligned} S^2 &= \frac{1}{n-1} \left[\Sigma d^2 - \frac{(\Sigma d)^2}{n} \right] = \frac{1}{8-1} \left[252 - \frac{(-38)^2}{8} \right] \\ &= \frac{1}{7} \left[252 - \left(\frac{1444}{8} \right) \right] = \frac{1}{7} [252 - 180.5] = \frac{71.5}{7} \\ S^2 &= 10.21 \end{aligned}$$

$$\therefore t = \frac{-4.75}{\sqrt{\frac{10.21}{7}}} = \frac{-4.75}{1.21} = -3.93$$

$$\begin{aligned} |t|_{\text{cal}} &= 3.93 & t_{\text{cal}} &= 3.93 \\ t_{\text{tab}} &= 1.90 \end{aligned}$$

(iii) **Decision:** Since $|t_{\text{cal}}| > t_{\text{tab}}$, we reject the null hypothesis at 5% level of significance and hence conclude that Meditation reduces BP level.

7. (c) Null hypothesis (H_0) : The attributes ‘liking for the new car model’ an ‘age groups’ are independent i.e. the car model appeals equally to the persons of all age groups.

Test statistic: Under H_0 , the expected frequencies are obtained as follows:

Computation of Expected Frequencies

	Age Group			
	Below 25	25–50	Above 50	Total
Liked the Car	$\frac{100 \times 100}{200} = 50$	$\frac{100 \times 50}{200} = 25$	$\frac{100 \times 50}{200} = 25$	100
Disliked the Car	$\frac{100 \times 100}{200} = 50$	$\frac{100 \times 50}{200} = 25$	$\frac{100 \times 50}{200} = 25$	100
Total	100	50	50	200

Computation for the Value of χ^2

O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
45	50	-5	25	0.5
30	25	5	25	1
25	25	0	0	0
55	50	5	25	0.5
20	25	-5	25	1
25	25	0	0	0
Total				3

$$\chi^2_{\text{cal}} = 3, \quad \chi^2_{\text{tab}} = 5.991$$

Conclusion: Since $\chi^2_{\text{cal}} < \chi^2_{\text{tab}}$, we accept the null hypothesis at 5% level of significance and hence conclude that the attributes ‘liking for the new car model’ and ‘age groups’ are independent.

8. (a) To test the validity of the claim or assumption about the population parameter, a sample is drawn from the population and analysed. The results of the analysis are used to decide whether the claim is true or not. The steps of general procedure for any hypothesis testing are summarized below:

Step 1: State the Null Hypothesis (H_0) and Alternative Hypothesis (H_1) The **null hypothesis** H_0 (read as H_0 sub-zero) represents the claim or statement made about the value or range of values of the population parameter. The capital letter H stands for hypothesis and the subscript 'zero' implies 'no difference' between sample statistic and the parameter value. Thus hypothesis testing requires that the null hypothesis be considered *true (status quo or no difference)* until it is proved false on the basis of results observed from the sample data. The null hypothesis is always expressed in the form of mathematical statement which includes the sign (\leq , $=$, \geq) making a claim regarding the specific value of the population parameter. That is:

$$H_0 : \mu (\leq, =, \geq) \mu_0$$

where μ is population mean and μ_0 represents a hypothesized value of μ . Only one sign out of \leq , $=$ and \geq will appear at a time when stating the null hypothesis.

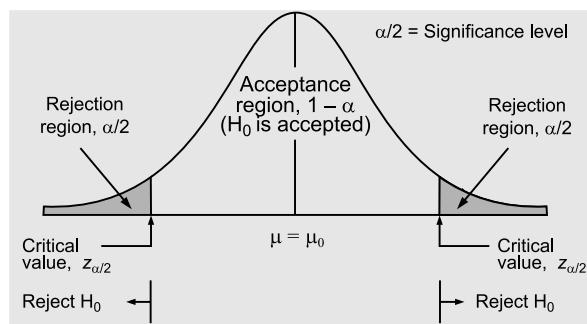
An **alternative hypothesis**, H_1 , is the counter claim (statement) made against the value of the particular population parameter. That is, an alternative hypothesis must be true when the null hypothesis is found to be false. In other words, the alternative hypothesis states that specific population parameter value is not equal to the value stated in the null hypothesis and is written as:

$$H_1 : \mu \neq \mu_0$$

Consequently $H_1 : \mu < \mu_0$ or $H_1 : \mu > \mu_0$

Step 2: State the Level of Significance, α (alpha): The level of significance, usually denoted by α (alpha), is specified before the samples are drawn, so that the results obtained should not influence the choice of the decision-maker. It is specified in terms of the probability of null hypothesis H_0 being wrong. In other words, the level of significance defines the likelihood of rejecting a null hypothesis when it is true, i.e. it is *the risk a decision-maker takes of rejecting the null hypothesis when it is really true*. The guide provided by the statistical theory is that this probability must be 'small'. Traditionally $\alpha = 0.05$ is selected for consumer research projects, $\alpha = 0.01$ for quality assurance and $\alpha = 0.10$ for political polling.

Step 3: Establish Critical or Rejection Region: The area under the sampling distribution curve of the test statistic is divided into two mutually exclusive regions (areas). These regions are called the *acceptance region* and the *rejection (or critical) region*.



Step 4: Select the Suitable Test of Significance or Test Statistic: The tests of significance or test statistic are classified into two categories: *parametric and nonparametric tests*. Parametric tests are more powerful because their data are derived from interval and ratio measurements. Nonparametric tests are used to test hypotheses with nominal and ordinal data. Parametric techniques are the tests of choice provided certain assumptions are met.

Nonparametric tests have few assumptions and do not specify normally distributed populations or homogeneity of variance.

Selection of a test. For choosing a particular test of significance following three factors are considered:

- (a) Whether the test involves one sample, two samples, or k samples?
- (b) Whether two or more samples used are independent or related?
- (c) Is the measurement scale nominal, ordinal, interval, or ratio?

One-sample tests are used for single sample and to test the hypothesis that it comes from a specified population.

The value of test statistic is calculated from the distribution of sample statistic by using the following formula

$$\text{Test statistic} = \frac{\text{Value of sample statistic} - \text{Value of hypothesized population parameter}}{\text{Standard error of the sample statistic}}$$

The choice of a probability distribution of a sample statistic is guided by the sample size n and the value of population standard deviation σ .

Step 5: Formulate a Decision Rule to Accept Null Hypothesis: Compare the calculated value of the test statistic with the critical value (also called *standard table value* of test statistic). The decision rules for null hypothesis are as follows:

- Accept H_0 if the test statistic value falls within the area of acceptance.
- Reject otherwise

In other words, if the calculated absolute value of a test statistic is less than or equal to its critical (or table) value, then accept the null hypothesis, otherwise reject it.

8. (b) H_0 : There is no significant difference in the proportion of defectives i.e.

$$H_0 : P_1 = P_2$$

$$H_1 : P_1 \neq P_2$$

Given:

$$n_1 = 80, n_2 = 130$$

$$p_1 = \frac{3}{80} = 0.0375$$

$$p_2 = \frac{2}{130} = 0.015$$

$$\text{Test Statistic: } z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

$$\begin{aligned} \hat{P} &= \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{80(0.0375) + 130(0.015)}{80 + 130} \\ &= \frac{3 + 1.95}{210} = \frac{4.95}{210} = 0.024 \end{aligned}$$

$$\hat{Q} = 1 - \hat{P} = 1 - 0.024 = 0.976$$

$$\begin{aligned} z &= \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.0375 - 0.015}{\sqrt{(0.024)(0.976)\left(\frac{1}{80} + \frac{1}{130}\right)}} \\ &= \frac{0.0225}{\sqrt{(0.023)(0.0195)}} = \frac{0.0225}{\sqrt{0.0004}} \end{aligned}$$

$$z = \frac{0.0225}{0.02} = 1.125$$

$$z_{\text{cal}} = 1.125$$

$$z_{\text{tab}} = 1.96$$

Since $z_{\text{cal}} < z_{\text{tab}}$, we accept the null hypothesis at 5% level of significance and hence conclude that there is no significant difference in the proportion of defectives.

8. (c) $H_0 : \mu_1 = \mu_2 = \mu_3$

i.e. there is no significant difference in the sales of the three salesman

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$$

A	B	C	A^2	B^2	C^2
300	600	700	90,000	3,60,000	4,90,000
400	300	300	1,60,000	90,000	90,000
300	300	400	90,000	90,000	1,60,000
500	400	600	2,50,000	1,60,000	3,60,000
-	-	500	-	-	2,50,000
1500	1600	2500	5,90,000	7,00,000	13,50,000

$$N = 13$$

$$\text{Grand total (G)} = 1500 + 1600 + 2500 = 5600$$

$$\text{Correction factor (CF)} = \frac{G^2}{N} = \frac{(5600)^2}{13} = \frac{3,13,60,000}{13} = 2412307.692$$

$$\begin{aligned} TSS &= 590000 + 700000 + 1350000 - 2412307.692 \\ &= 227692.308 \end{aligned}$$

$$\begin{aligned} SSR &= \frac{(1500)^2}{4} + \frac{(1600)^2}{4} + \frac{(2500)^2}{5} - 2412307.692 \\ &= 562500 + 640000 + 1250000 - 2412307.692 \\ &= 40192.308 \end{aligned}$$

$$\begin{aligned} SSE &= TSS - SSR \\ &= 227692.308 - 40192.308 = 187500 \end{aligned}$$

Anova

<i>Source of variation</i>	<i>Degree of freedom</i>	<i>Sum of Squares</i>	<i>Mean sum of Squares</i>	<i>F-ratio</i>
Rows	$(3 - 1) = 2$	40192.308	20096.154	$1.07 \sim F_{(2,10)}$
Error	$(12 - 2) = 10$	187500	18750	
Total	$(13 - 1) = 12$	227692.308	-	

$$F_{\text{cal}} = 1.07, \quad F_{\text{tab}} = 4.10$$

Since $F_{\text{cal}} < F_{\text{tab}}$, we accept the null hypothesis at 5% level of significance and hence conclude that there is no significant difference in the sales of the three salesman.

SOLUTION TO MODEL QUESTION PAPER-II

1. (a) Individuals, focus groups, and/or panels of respondents specifically decided upon and set up by the investigator for data collection are examples of primary data sources. Any one or a combination of the following methods can be chosen to collect primary data:
- (i) Direct personal observations
 - (ii) Direct or indirect oral interviews
 - (iii) Administrating questionnaires

Secondary data refer to those data which have been collected earlier for some purpose other than the analysis currently being undertaken. Besides newspapers and business magazines, other sources of such data are as follows:

External secondary data sources Government publications, which include

- (i) The National Accounts Statistics, published by the Central Statistical Organization (CSO). It contains estimates of national income for several years, growth rate, and rate on major economic activities such as agriculture, industry, trade, transport, and so on;
- (ii) Wholesale Price Index, published by the office of the Economic Advisor, Ministry of Commerce and Industry;
- (iii) Consumer Price Index;
- (iv) Reserve Bank of India Bulletins;
- (v) Economic Survey.

Non-Government publications include publications of various industrial and trade associations such as

- (i) The Indian Cotton Mills Association
- (ii) The various Chambers of Commerce
- (iii) The Bombay Stock Exchange, which publishes a directory containing financial accounts, key profitability and other relevant data.

Various syndicate services such as Operations Research Group (ORG). The Indian Market Research Bureau (IMRB) also collects and tabulates abundant marketing information to suit the requirements of individual firms, making the same available at regular intervals.

International organizations which publish data are:

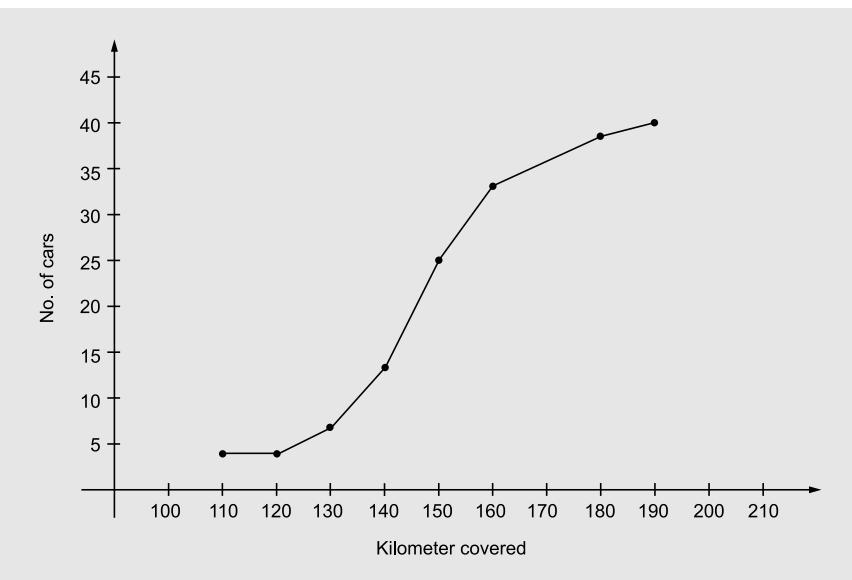
- (i) The International Labour Organization (ILO)—which publishes data on the total and active population, employment, unemployment, wages, and consumer prices.
- (ii) The Organization for Economic Cooperation and Development (OECD)—which publishes data on foreign trade, industry, food, transport, and science and technology.
- (iii) The International Monetary Fund (IMF)—which publishes reports on national and international foreign exchange regulations and other trade barriers, foreign trade, and economic developments.

$$\begin{aligned}1. (b) \quad I &= \frac{10}{100}x, \quad I = \frac{12}{100}y, \quad I = \frac{15}{100}z \\ \Rightarrow \quad x &= \frac{100}{10}I, \quad y = \frac{100}{12}I, \quad z = \frac{100}{15}I \\ x + y + z &= 100I\left(\frac{1}{10} + \frac{1}{12} + \frac{1}{15}\right) = 100I\left(\frac{6+5+4}{60}\right) \\ &= \frac{100I}{60} \left(\frac{5}{60} \right) \text{ {Due to average}} \\ &= \frac{50}{6}I = \frac{25}{3}I\end{aligned}$$

$$\begin{aligned}x + y + z &= \frac{25}{3} I \times 100 \quad \{\text{percentage}\} \\&= \frac{2500}{3} I \%\end{aligned}$$

1. (c)

Kilometers covered	No. of cars <i>f</i>	<i>cf</i>
100–110	4	4
110–120	0	4
120–130	3	7
130–140	7	14
140–150	11	25
150–160	8	33
160–170	5	38
170–180	0	38
180–190	2	40
$\sum N = 40$		

To find Q_1 :

$$N = 40, \frac{N}{4} = 10, h = 10, l = 130, f = 7, C = 7$$

$$\begin{aligned}Q_1 &= l + \left[\frac{h}{f} \left(\frac{N}{4} - C \right) \right] = 130 + \left[\frac{10}{7} (10 - 7) \right] \\&= 130 + [1.43(3)] = 130 + [4.29] = 134.29\end{aligned}$$

To find Q_2 :

$$N = 40, \frac{N}{2} = 20, l = 140, h = 10, f = 11, C = 14$$

$$\begin{aligned}Q_2 &= l + \left[\frac{h}{f} \left(\frac{N}{2} - C \right) \right] = 140 + \left[\frac{10}{11} (20 - 14) \right] \\&= 140 + [0.91(6)] = 140 + 5.46 = 145.46\end{aligned}$$

To find Q_3 :

$$N = 40, \frac{3N}{4} = \frac{3(40)}{4} = \frac{120}{4} = 30; l = 150, h = 10, f = 8, C = 25$$

$$\begin{aligned}
 Q_3 &= l + \left[\frac{h}{f} \left(\frac{3N}{4} - C \right) \right] = 150 + \left[\frac{10}{8} (30 - 25) \right] \\
 &= 150 + [1.25(5)] = 150 + (6.25) = 156.25
 \end{aligned}$$

To find P_{75} :

$$\text{Percentile } 75^{\text{th}} \text{ item} = \frac{75N}{100} = \frac{75(40)}{100} = \frac{3000}{100} = 30$$

The cumulative frequency just greater than 30 is 33. Hence the corresponding class 150–160 contains P_{75}

$$\begin{aligned}
 \therefore P_{75} &= 150 + \left[\frac{30 - 25}{8} \right] \times 10 = 150 + \left[\frac{5}{8} \right] \times 10 \\
 &= 150 + (0.625 \times 10) = 150 + 6.25 = 156.25
 \end{aligned}$$

2. (a) Definition of index numbers can be classified into the following three broad categories:

A measure of change: It is a numerical value characterizing the change in complex economic phenomena over a period of time or space.

—Maslow

A device to measure change: Index numbers are devices measuring differences in the magnitude of a group of related variables. —Croxton and Cowden

A series representing the process of change: Index numbers are series of numbers by which changes in the magnitude of a phenomenon are measured from time to time or place to place. —Horace Secris

Uses of Index Numbers: According to G. Simpson and F. Kafka ‘Today Index numbers are one of the most widely used statistical tools. They are used to feel the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies’. Other important uses of index number can be summarized as follows:

1. **Index numbers act as economic barometers:** A barometer is an instrument that is used to measure atmospheric pressure. Index numbers are used to feel the pressure of the economic and business behaviour, as well as to measure ups and downs in the general economic condition of a country. For example, the composite index number of indexes of prices, industrial output, foreign exchange reserves, and bank deposits, could act as an economic barometer.
2. **Index numbers help in policy formulation:** Many aspects of economic activity are related to price movements. The price indexes can be used as indicators of change in various segments of the economy. For example, by examining the price indexes of different segments of a firm’s operations, the management can assess the impact of price changes and accordingly take some remedial and/or preventive actions.
3. **Index numbers reveal trends and tendencies:** An index number is defined as a relative measure describing the average change in the level of a phenomenon between the current period and a base period. This property of the index number can be used to reflect typical patterns of change in the level of a phenomenon. For example, by examining the index number of industrial production, agricultural production, imports, exports, and wholesale and retail prices for the last 8–10 years, we can draw the trend of the phenomenon under study and also draw conclusions as to how much change has taken place due to the various factors.
4. **Index numbers help to measure purchasing power:** In general, the purchasing power is not associated with a particular individual; rather it is related to an entire class or group. Furthermore, it is not associated with the cost of a single item, because individuals purchase many different items in order to live. Consequently, earnings of a group of people or class must be adjusted with a price index that provides an overall view of the purchasing power for the group.

5. **Index numbers help in deflating various values:** When real rupee value is computed, the base period is earlier than the given years for which this value is being determined. Thus the adjustment of current rupee value to real terms is referred to as *deflating a value series* because prices typically increase over time.

The price index number is helpful in deflating the national income to remove the effect of inflation over a long term, so that we may understand whether there is any change in the real income of the people or not. The retail price index is often used to compute real changes in earnings and expenditure as it compares the purchasing power of money at different points in time. It is generally accepted as a standard measure of inflation even though calculated from a restricted basket of goods.

2. (b) Given: Before settlement:

$$\bar{x} = 800, \sigma = 100 \quad CV \rightarrow \text{Coefficient of variation}$$

$$\therefore CV_1 = \frac{\sigma}{\bar{x}} \times 100 = \frac{100}{800} \times 100 = 0.125 \times 100 = 12.5$$

After settlement:

$$\bar{x} = 1200, \sigma = 150$$

$$CV_2 = \frac{\sigma}{\bar{x}} \times 100 = \frac{150}{1200} \times 100$$

$$CV_2 = 0.125 \times 100 = 12.5$$

Since CV_1 and CV_2 are same, we conclude that the wages are uniform before and after settlement.

2. (c) To calculate average output:

Output (in units)	Frequency <i>f</i>	Mid-point <i>x</i>	<i>fx</i>
70–74	3	72	216
75–79	5	77	385
80–84	15	82	1230
85–89	12	87	1044
90–94	7	92	644
95–99	6	97	582
100–104	2	102	204
	50		4305

$$\bar{X} = \frac{\sum fx}{\sum f} = \frac{4305}{50} = 86.1$$

∴ Average output = 86.1

To calculate the average bonus:

Bonus (Rs.) <i>X</i>	Frequency <i>f</i>	<i>fx</i>
40	3	120
45	5	225
50	15	750
60	12	720
70	7	490
80	6	480
100	2	200
	50	2985

$$\bar{X} = \frac{\sum fx}{\sum f} = \frac{2985}{50} = 59.7$$

∴ The average bonus = 59.7

3. (a) Correlation analysis contributes to the understanding of economic behaviour, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilizing forces may become effective.

The fundamental aim of regression analysis is to determine a regression equation (line) that makes sense and fits the representative data such that the error of variance is as small as possible. This implies that the regression equation should adequately be used for prediction. J. R. Stockton stated that

The device used for estimating the values of one variable from the value of the other consists of a line through the points, drawn in such a manner as to represent the average relationship between the two variables. Such a line is called line of regression.

The two variables x and y which are correlated can be expressed in terms of each other in the form of straight line equations called *regression equations*. Such lines should be able to provide the best fit of sample data to the population data. The algebraic expression of regression lines is written as:

The regression equation of y on x

$$y = a + bx$$

is used for estimating the value of y for given values of x .

Regression equation of x on y

$$x = c + dy$$

$$N = 3$$

is used for estimating the value of x for given values of y .

3. (b) Average rate of growth:

$$\begin{aligned} GM &= A \log \left[\frac{\Sigma \log x}{N} \right] = A \log \left[\frac{2.68}{3} \right] \\ &= A \log [0.89] = 7.76\% \end{aligned}$$

x	$\log x$
5	0.7
8	0.9
12	1.08
	2.68

3. (c) Karl Pearson's co-efficient of correlation

Trunk height x	Diameter y	xy	x^2	y^2
35	8	280	1225	64
49	9	441	2401	81
27	7	189	729	49
33	6	198	1089	36
60	13	780	3600	169
21	7	147	441	49
45	11	495	2025	121
51	12	612	2601	144
321	73	3142	14111	713

$$\begin{aligned} r &= \frac{N \Sigma xy - \Sigma x \Sigma y}{\sqrt{N \Sigma x^2 - (\Sigma x)^2} \sqrt{N \Sigma y^2 - (\Sigma y)^2}} \\ &= \frac{8(3142) - (321)(73)}{\sqrt{8(14111) - (321)^2} \sqrt{8(713) - (73)^2}} \\ &= \frac{25136 - 23433}{\sqrt{112888 - 103041} \sqrt{5704 - 5329}} \\ &= \frac{1703}{(99.23)(19.36)} = \frac{1703}{1921.0928} \\ r &= 0.8865 \end{aligned}$$

4. (a) Although frequency distributions and corresponding graphical representations make raw data more meaningful, yet they fail to identify three major properties that describe a set of quantitative data. These three major properties are:

1. The numerical value of an observation (also called *central value*) around which most numerical values of other observations in the data set show a tendency to cluster or group, called the *central tendency*.
2. The extent to which numerical values are dispersed around the central value, called *variation*.
3. The extent of departure of numerical values from symmetrical (normal) distribution around the central value, called *skewness*.

These three properties—*central tendency*, *variation*, and *shape* of the frequency distribution—may be used to extract and summarize major features of the data set by the application of certain statistical methods called *descriptive measures* or *summary measures*. There are three types of summary measures:

1. Measures of central tendency
2. Measures of dispersion or variation
3. Measure of symmetry—skewness

Just as central tendency can be measured by a number in the form of an average, the amount of variation (dispersion, spread, or scatter) among the values in the data set can also be measured. The measures of central tendency describe that the major part of values in the data set appears to concentrate (cluster) around a central value called *average* with the remaining values scattered (spread or distributed) on either sides of that value. But these measures do not reveal how these values are dispersed (spread or scattered) on each side of the central value. The dispersion of values is indicated by the extent to which these values tend to spread over an interval rather than cluster closely around an average.

The statistical techniques to measure such dispersion are of two types:

- (a) Techniques that are used to measure the extent of variation or the deviation (also called degree of variation) of each value in the data set from a measure of central tendency, usually the mean or median. Such statistical techniques are called *measures of dispersion* (or *variation*).
- (b) Techniques that are used to measure the direction (away from uniformity or symmetry) of variation in the distribution of values in the data set. Such statistical techniques are called *measures of skewness*.

To measure the dispersion, understand it, and identify its causes is very important in statistical inference (estimation of parameter, hypothesis testing, forecasting, and so on). A small dispersion among values in the data set indicates that data are clustered closely around the mean. The mean is therefore considered representative of the data, i.e. mean is a reliable average. Conversely, a large dispersion among values in the data set indicates that the mean is not reliable, i.e. it is not representative of the data.

The degree of skewness in a distribution can be measured both in the *absolute* and *relative* sense. For an asymmetrical distribution, the distance between mean and mode may be used to measure the degree of skewness because the mean is equal to mode in a symmetrical distribution. Thus,

$$\begin{aligned} \text{Absolute } S_k &= \text{Mean} - \text{Mode} \\ &= Q_3 + Q_1 - 2 \text{ Median} \quad (\text{if measured in terms of quartiles}). \end{aligned}$$

For a positively skewed distribution, Mean > Mode and therefore S_k is a positive value, otherwise it is a negative value. This difference is taken to measure the degree of skewness because in an asymmetrical distribution, mean moves away from the mode. Larger the difference between mean and mode, whether positive or negative, more is the asymmetrical distribution or skewness. This difference, however, may not be desirable for the following reasons:

- (i) The difference between mean and mode is expressed in the same units as the distribution and therefore cannot be used for comparing skewness of two or more distributions having different units of measurement.
- (ii) The difference between mean and mode may be large in one distribution and small in another, although the shape of their frequency curves is the same.

In order to overcome these two shortcomings and to make valid comparisons between skewness of two or more distributions, the absolute difference has to be expressed in relation to the standard deviation—a measure of dispersion. Since we want to express any measure of skewness as a pure (relative) number, therefore this distance is expressed in terms of the unit of measurement in units of the standard deviation.

4. (b) $P(\text{ATM is used}) = P(A) = \frac{30}{60} = \frac{1}{2}$

$$P(\text{ATM is not used}) = 1 - P(A) = 1 - \frac{1}{2} = \frac{1}{2}$$

$$P(A) \text{ in 10 minutes} = 10 \times \frac{1}{2} = 5$$

$$(i) P(\text{no body uses the machine}) \text{ in 10 minutes} = 5 \times 10 = 0$$

$$(ii) P(3 \text{ people uses the machine}) \text{ in 10 minutes} = \frac{6}{60} = \frac{1}{10}$$

$$1 \text{ person} = 2 \text{ min}$$

$$3 \text{ persons} = 2 \times 3 = 6 \text{ min}$$

4. (c)	<i>Production</i> <i>X</i>	<i>Capacity utilization (in %)</i> <i>Y</i>
	$\bar{X} = 35.6$	$\bar{Y} = 84.8$
	$\sigma_X = 10.5$	$\sigma_Y = 8.5$

$$r = 0.62$$

Regression equation showing the capacity utilization on production: i.e. Y on X
 Y on X :

$$\begin{aligned} Y - \bar{Y} &= b_{YX}(X - \bar{X}) & b_{YX} &= r \cdot \frac{\sigma_Y}{\sigma_X} \\ Y - 84.8 &= 0.5(X - 35.6) & &= \frac{(0.62)(8.5)}{10.5} \\ Y - 84.8 &= 0.5 \times -17.8 & &= 0.5 \\ Y &= 0.5X - 17.8 + 84.8 & &= 0.5 \\ Y &= 0.5X + 67 \end{aligned}$$

To estimate the production when capacity utilization is 70%:

X on Y :

$$\begin{aligned} X - \bar{X} &= b_{XY}(Y - \bar{Y}) & b_{XY} &= r \cdot \frac{\sigma_X}{\sigma_Y} \\ X - 35.6 &= 0.77(Y - 84.8) & &= \frac{(0.62)(10.5)}{8.5} \\ X - 35.6 &= 0.77Y - 65.3 & &= \frac{6.51}{8.5} = 0.77 \\ X &= 0.77Y - 65.3 + 35.6 & & \\ X &= 0.77Y - 29.7 & & \end{aligned}$$

when $Y = 70$

$$\Rightarrow X = 0.77(70) - 29.7 = 53.9 - 29.7$$

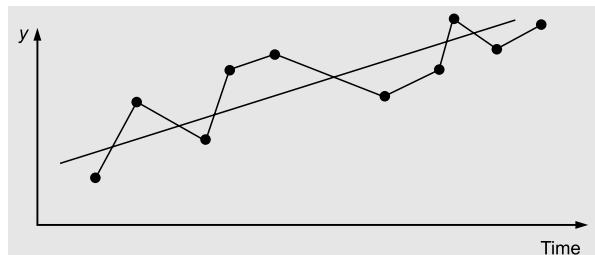
$$X = 24.2$$

5. (a) A time series is a set of numerical values of some variable obtained at regular period over time. The series is usually tabulated or graphed in a manner that readily conveys the behaviour of the variable under study.

The **time-series** data contain four components: *trend*, *cyclical*, *seasonality* and *irregularity*. Not all time-series have all these components. Figure 16.4 shows the effects of these time-series components over a period of time.

Trend: Sometimes a time-series displays a steady tendency of either upward or downward movement in the average (or mean) value of the forecast variable y over time. Such a tendency is called a trend. When observations are plotted against time, a straight line describes the increase or decrease in the time series over a period of time.

Cycles: An upward and downward movement in the variable value about the trend time over a time period are called cycles. A business cycle may vary in length, usually more than a year but less than 5 to 7 years. The movement is through four phases: from *peak* (prosperity) to *contradiction* (recession) to *trough* (depression) to *expansion* (recovery or growth) as shown in figure.



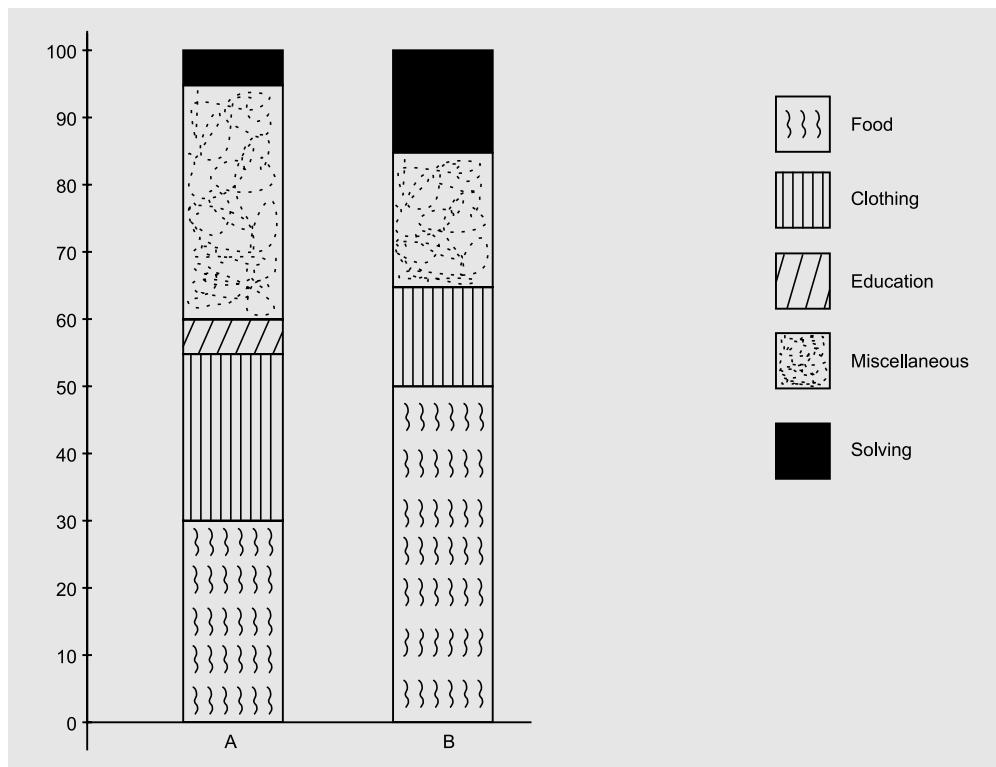
Time-series Effects

Seasonal: It is a special case of a cycle component of time series in which fluctuations are repeated usually within a year (e.g. daily, weekly, monthly, quarterly) with a high degree of regularity. For example, average sales for a retail store may increase greatly during festival seasons.

5. (b) Percentage sub-divided bar diagram"

Calculation: Expressed in percentage

Item of Expenditure	Family A	Family B
Food	30	50
Clothing	25	20
Education	5	17
Miscellaneous	38	23
Saving	2	-10
	100	100



5. (c) Laspeyre's, Paasche's and Fisher's index numbers

Commodity	p_0	p_1	q_0	q_1	$q_1 p_0$	$q_0 p_0$	$q_1 p_1$	$q_0 p_1$
A	20	40	8	6	120	160	240	320
B	50	60	10	5	250	500	300	600
C	40	50	15	15	600	600	750	750
D	20	20	20	25	500	400	500	400
					1470	1660	1790	2070

Laspeyre's index numbers:

$$p_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{2070}{1660} \times 100 = 1.25 \times 100 = 125$$

Paasche's index number:

$$p_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{1790}{1470} \times 100 = 1.22 \times 100 = 122$$

Fisher's index number:

$$\begin{aligned} p_{01} &= \sqrt{L \times P} = \sqrt{(125)(122)} && L \rightarrow \text{Laspeyre's index} \\ &= \sqrt{15250} = 123.5 && P \rightarrow \text{Paasche's index} \end{aligned}$$

6. (a) Poisson distribution is named after the French mathematician S. Poisson. The Poisson process measures the number of occurrences of a particular outcome of a discrete random variable in a *predetermined time interval, space, or volume*, for which an *average number* of occurrences of the outcome is known or can be determined. In the Poisson process, the random variable values need counting. Such a count might be (i) number of telephone calls per hour coming into the switchboard, (ii) number of fatal traffic accidents per week in a city/state. The Poisson probability distribution provides a simple, easy-to compute and accurate approximation to a binomial distribution when the probability of success, p is very small and n is large, so that $\mu = np$ is small, preferably $np > 7$. It is often called the '*law of improbable*' events meaning that the probability, p , of a particular event's happening is very small. **Poisson distribution** occurs in business situations in which there are a few successes against a large number of failures or vice-versa (i.e. few successes in an interval) and has single independent events that are mutually exclusive. Because of this, the probability of success, p is very small in relation to the number of trials n , so we consider only the probability of success.

Since Poisson probability distribution is specified by a process rate λ and the time period t , its mean and variance are identical and are expressed in terms of the parameters: n and p .

Poisson distribution is defined by the parameter λ and is positively skewed and leptokurtic. This implies that there is a possibility of infinitely large number of occurrences in a particular time interval, even though the average rate of occurrences is very small. However, as $\lambda \rightarrow \infty$, the distribution tends to be symmetrical and mesokurtic.

It is very *rare for more than one event to occur* during a short interval of time. The shorter the duration of interval, the occurrence of two or more events also becomes rare. The probability that exactly one event will occur in such an interval is approximately λ times its duration.

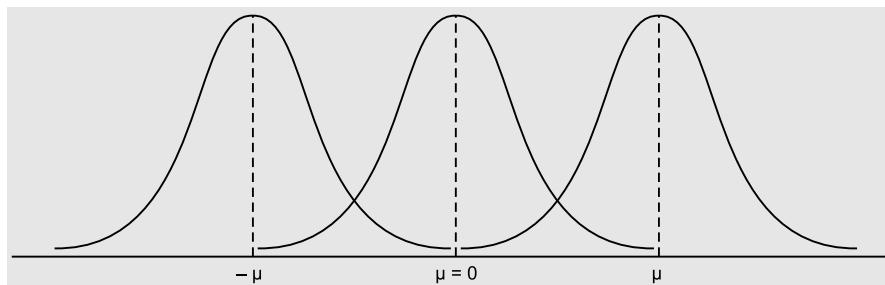
If λ is not an integer and $m = [\lambda]$, the largest integer contained in it, then m is the unique mode of the distribution. But if λ is an integer, the distribution would be bimodal.

The typical application of Poisson distribution is for analysing queuing (or waiting line) problems in which arriving customers during an interval of time arrive independently and the number of arrivals depends on the length of the

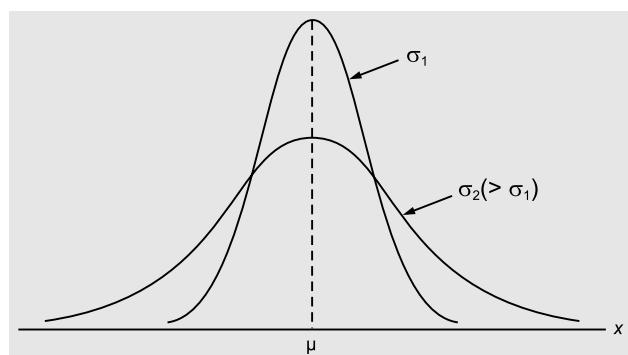
time interval. While applying Poisson distribution if we consider a time period of different length, the distribution of number of events remains Poisson with the mean proportional to the length of the time period.

If a random variable is discrete, then it is possible to assign a specific probability to each of its value and get the probability distribution for it. The sum of all the probabilities associated with the different values of the random variable is 1. However, not all experiments result in random variables that are discrete. Continuous random variables such as height, time, weight, monetary values, length of life of a particular product, etc. can take large number of observable values corresponding to points on a line interval much like the infinite number of grains of sand on a beach. The sum of probability to each of these infinitely large values is no longer sum to 1.

Characteristics of the Normal Probability Distribution: There is a family of normal distributions. Each normal distribution may have a different mean μ or standard deviation σ . A unique normal distribution may be defined by assigning specific values to the mean μ and standard deviation σ in the normal probability density function. Large value of σ reduce the height of the curve and increase the spread; small values of σ increase the height of the curve and reduce the spread. Shows three normal distributions with different values of the mean μ and a fixed standard deviation σ , while in normal distributions are shown with different values of the standard deviation σ and a fixed mean μ .



Normal Distributions with Different Mean Values But Fixed Standard Deviation



Normal Distributions with Fixed Mean and Variable Standard Deviation

From the above figure the following characteristics of a normal distribution and its density function may be derived:

- (i) For every pair of values of μ and σ , the curve of normal probability density function is bell shaped and symmetric.
- (ii) The normal curve is symmetrical around a vertical line erected at the mean μ with respect to the area under it, that is, fifty per cent of the area of the curve lies on both sides of the mean and reflect the mirror image of the shape of the curve on both sides of the mean μ . This implies that the probability of any individual outcome above or below the mean will be same. Thus, for any normal random variable x ,

$$P(x \leq \mu) = P(x \geq \mu) = 0.50$$

- (iii) Since the normal curve is symmetric, the mean, median, and mode for the normal distribution are equal because the highest value of the probability density function occurs when value of a random variable, $x = \mu$.
- (iv) The two tails of the normal curve extend to infinity in both directions and theoretically never touch the horizontal axis.
- (v) The mean of the normal distribution may be negative, zero, or positive as shown in figure.
- (vi) The mean μ determines the *central location* of the normal distribution, while standard deviation σ determines its *spread*. The larger the value of the standard deviation σ , the wider and flatter is the normal curve, thus showing more variability in the data, as shown in figure. Thus standard deviation σ determines the range of values that any random variable is likely to assume.
- (vii) The area under the normal curve represents probabilities for the normal random variable, and therefore, the total area under the curve for the normal probability distribution is 1.
6. (b) The first step in the analysis of variance is to partition the total variation in the sample data into the following two component variations in such a way that it is possible to estimate the contribution of factors that may cause variation.
1. The amount of variation *among the sample means* or the variation attributable to the difference among sample means. This variation is either on account of difference in treatment or due to element of chance. This difference is denoted by SSC or SSTR.
 2. The amount of variation *within the sample observations*. This difference is considered due to chance causes or experimental (random) errors. The difference in the values of various elements in a sample due to chance is called an estimate and is denoted by SSE.
- The observations in the sample data may be classified according to *one factor* (criterion) or *two factors* (criteria). The classifications according to one factor and two factors are called *one-way classification* and *two-way classification*, respectively. The calculations for total variation and its components may be carried out in each of the two-types of classifications by (i) *direct method*, (ii) *shortcut method*, and (iii) *coding method*.
6. (c) Linear trend:

Year <i>t</i>	Sales <i>Y</i>	$x = t - 2001$	X^2	XY
1998	550	-3	9	-1650
1999	560	-2	4	-1120
2000	555	-1	1	-555
2001	585	0	0	0
2002	540	1	1	540
2003	525	2	4	1050
2004	545	3	9	1635
2005	585	4	16	2340
$\Sigma Y = 4445$		$\Sigma X = 4$	$\Sigma X^2 = 44$	$\Sigma XY = 2240$

Normal equations:

$$\Sigma Y = na + b\Sigma X$$

$$4445 = 8a + 4b \quad \dots(1)$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

$$2240 = 4a + 44b \quad \dots(2)$$

Solving (1) and (2)

$$\begin{array}{ll}
 (1) & 8a + 4b = 4445 \\
 (2) \times 2 & 8a + 88b = 4480 \\
 & (-) \quad (-) \quad (-) \\
 & \hline
 & -84b = -35 \\
 b & = \frac{35}{84} \\
 b & = 0.42
 \end{array}$$

Substitute b in (1)

$$\begin{aligned}
 8a + 4(0.42) &= 4445 \\
 8a + 1.68 &= 4445 \\
 8a &= 4445 - 1.68 \\
 8a &= 4443.32 \\
 a &= \frac{4443.32}{8} \\
 a &= 555.415
 \end{aligned}$$

∴ Linear trend fitted to yearly values is:

$$\begin{aligned}
 Y &= a + bX \\
 \Rightarrow Y &= 555.415 + 0.42X
 \end{aligned}$$

Sales for the year 1997:

i.e. when $X = -4$

$$\begin{aligned}
 Y_{1997} &= 555.415 + 0.42(-4) \\
 Y_{1997} &= 555.415 + 0.42(-4) \\
 &= 555.415 - 1.68 \\
 Y &= 553.735 \text{ millions}
 \end{aligned}$$

Slope:

$$\begin{aligned}
 Y &= a + bX \\
 Y &= 555.415 + 0.42X
 \end{aligned}$$

Slope = 0.42

7. (a) Sampling methods compared to census provides an attractive means of learning about a population or process in terms of reduced cost, time and greater accuracy. The representation basis and the element selection techniques from the given population, classify several sampling methods into two categories as shown in the table.

Types of Sampling Methods

Element Selection	Representation Basis	
	Probability (Random)	Non-probability (Non-random)
• Unrestricted	Simple random sampling	Convenience sampling
• Restricted	Complex random sampling	Purposive sampling
	• Stratified sampling	• Quota sampling
	• Cluster sampling	• Judgement sampling
	• Systematic sampling	
	• Multi-stage sampling	

Probability Sampling Methods

Several probability sampling methods for selecting samples from a population or process are as follows:

Simple Random (Unrestricted) Sampling: In this method, every member (or element) of the population has an equal and independent chance of

being selected again and again when a sample is drawn from the population. To draw a random sample, we need a complete list of all elements in the population of interest so that each element can be identified by a distinct number. Such a list is called *frame for experiment*. The frame for experiment allows us to draw elements from the population by randomly generating the numbers of the elements to be included in the sample.

This method is suitable for sampling, as many statistical tests assume independence of sample elements. One disadvantage with this method is that all elements of the population have to be available for selection, which many a times is not possible.

Stratified Sampling: This method is useful when the population consists of a number of heterogeneous subpopulations and the elements within a given subpopulation are relatively homogeneous compared to the population as a whole. Thus, population is divided into mutually exclusive groups called *strata* that are relevant, appropriate and meaningful in the context of the study. A simple random sample, called a *sub-sample*, is then drawn from each *strata* or *group*, in proportion or a non-proportion to its size. As the name implies, a proportional sampling procedure requires that the number of elements in each stratum be in the same proportion as in the population. In non-proportional procedure, the number of elements in each stratum are disproportionate to the respective numbers in the population. The basis for forming the strata such as location, age, industry type, gross sales, or number of employees, is at the discretion of the investigator. Individual stratum samples are combined into one to obtain an overall sample for analysis.

This sampling procedure is more efficient than the simple random sampling procedure because, for the same sample size, we get more representativeness from each important segment of the population and obtain more valuable and differentiated information with respect to each strata.

Cluster Sampling: This method, sometimes known as *area sampling method*, has been devised to meet the problem of costs or inadequate sampling frames (a complete listing of all elements in the population so that each member can be identified by a distinct number). The entire population to be analysed is divided into smaller groups or chunks of elements and a sample of the desired number of areas selected by a simple random sampling method. Such groups are termed as *clusters*. The elements of a cluster are called *elementary units*. These clusters do not have much heterogeneity among the elements. A household where individuals live together is an example of a cluster.

If several groups with intragroup heterogeneity and intergroup homogeneity are found, then a random sampling of the clusters or groups can be done with information gathered from each of the elements in the randomly chosen clusters. Cluster samples offer more heterogeneity within groups and more homogeneity among groups—the reverse of what we find in stratified random sampling, where there is homogeneity within each group and heterogeneity across groups. For instance, committees formed from various departments in an organization to offer inputs to make decisions on product development, budget allocations, marketing strategies, etc. are examples of different clusters. Each of these clusters or groups contains a heterogeneous collection of members with different interests, orientations, values, philosophy, and vested interests. Based on individual and combined perceptions, it is possible to make final decision on strategic moves for the organization.

In summary, cluster sampling involves preparing only a list of clusters instead of a list of individual elements. For examples, (i) residential blocks (colonies) are commonly used to cluster in surveys that require door-to-door interviews, (ii) airlines sometimes select randomly a set of flights to distribute questionnaire to every passenger on those flights to measure customer satisfaction. In this situation, each flight is a cluster. It is much easier for the airline to choose a random sample of flights than to identify and locate a random sample of individual passengers to distribute questionnaire.

Multistage Sampling: This method of sampling is useful when the population is very widely spread and random sampling is not possible. The researcher might stratify the population in different regions of the country, then stratify by urban and rural and then choose a random sample of communities within these strata. These communities are then divided into city areas as clusters and randomly consider some of these for study. Each element in the selected cluster may be contacted for desired information.

For example, for the purpose of a national pre-election opinion poll, the *first stage* would be to choose as a sample a specific state (region). The size of the sample, that is the number of interviews, from each region would be determined by the relative populations in each region. In the *second stage*, a limited number of towns/cities in each of the regions would be selected, and then in the *third stage*, within the selected towns/cities, a sample of respondents could be drawn from the electoral roll of the town/city selected at the second stage.

The essence of this type of sampling is that a subsample is taken from successive groups or strata. The selection of the sampling units at each stage may be achieved with or without stratification. For example, at the second stage when the sample of towns/cities is being drawn, it is customary to classify all the urban areas in the region in such a way that the elements (towns/cities) of the population in those areas are given equal chances of inclusion.

Systematic Sampling: This procedure is useful when elements of the population are already physically arranged in some order, such as an alphabetized list of people with driving licenses, list of bank customers by account numbers. In these cases one element is chosen at random from first k element and then every k th element (member) is included in the sample. The value k is called the *sampling interval*. For example, suppose a sample size of 50 is desired from a population consisting of 100 accounts receivable. The sampling interval is $k = N/n = 1000/50 = 20$. Thus a sample of 50 accounts is identified by moving systematically through the population and identifying every 20th account after the first randomly selected account number.

Non-Random Sampling Methods

Several non-random sampling methods for selecting samples from a population or process are as follows:

Convenience Sampling: In this procedure, units to be included in the sample are selected at the convenience of the investigator rather than by any prespecified or known probabilities of being selected. For example, a student for his project on 'food habits among adults' may use his own friends in the college to constitute a sample simply because they are readily available and will participate for little or no cost. Other examples are, public opinion surveys conducted by any TV channel near the railway station; bus stop, or in a market.

Convenience samples are easy for collecting data on a particular issue. However, it is not possible to evaluate its representativeness of the population and hence precautions should be taken in interpreting the results of convenient samples that are used to make inferences about a population.

Purposive Sampling: Instead of obtaining information from those who are most conveniently available, it sometimes becomes necessary to obtain information from specific targets—respondents who will be able to provide the desired information either because they are the only ones who can give the desired information or because they satisfy to some criteria set by researcher.

Judgement Sampling: Judgement sampling involves the selection of respondents who are in the best position to provide the desired information. The judgment sampling is used when a limited number of respondents have the information that is needed. In such cases, any type of probability sampling across a cross section of respondents is purposeless and not useful. This sampling method may curtail the generalizability of the findings due to the fact that we are using a sample of respondents who are conveniently available to us. It is the only viable sampling method for obtaining the type of information that is required from very specific section of respondents who possess the knowledge and can give the desired information.

However, the validity of the sample results depend on the proper judgment of the investigator in choosing the sample. Great precaution is needed in drawing conclusions based on judgment samples to make inferences about a population.

Quota Sampling: Quota Sampling is a form of proportionate stratified sampling in which a predetermined proportion of elements are sampled from different groups in the population, but on convenience basis. In other words, in quota sampling the selection of respondents lies with the investigator, although in making such selection he/she must ensure that each respondent satisfies certain criteria which is essential for the study.

7. (b) $n = 25$

$$\bar{x} = 30,000$$

$$S = 10,000$$

$$\text{Interval estimate: } \bar{x} \pm t_{0.05}(n-1) \cdot \frac{S}{\sqrt{n-1}}$$

$$\Rightarrow (30,000 - 1.711) \cdot \frac{10,000}{\sqrt{24}} = (29998.3) \frac{10000}{4.9} \\ = (29998.3) (2040.8) = 6,12,20,530.64$$

$$\bar{x} + t_{0.05}(n-1) \cdot \frac{S}{\sqrt{n-1}}$$

$$\Rightarrow (30,000 + 1.711) \cdot \frac{10000}{\sqrt{24}} \\ (30001.7) (2040.8) = 61227469.4$$

∴ Interval estimate:

$$61220530.64 < \theta < 61227469.4$$

7. (c) H_0 : There is no significant difference between the average sales of the two salesmen

$$\begin{aligned} \text{S.E. } (\bar{X}_1 - \bar{X}_2) &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(70)^2}{20} + \frac{(60)^2}{22}} \\ &= \sqrt{\frac{4900}{20} + \frac{3600}{22}} = 20.2 \end{aligned}$$

$$\frac{\text{Difference}}{\text{S.E.}} = \frac{800-780}{20.2} = 0.99$$

$$\begin{aligned} \sigma_1 &= 70 \\ \sigma_2 &= 60 \\ n_1 &= 20 \\ n_2 &= 22 \\ \bar{X}_1 &= 800 \\ \bar{X}_2 &= 780 \end{aligned}$$

Since the difference is less than 2.58 S.E. (1% level of significance) the hypothesis is accepted i.e. there is no significant difference between the average sales of the two salesmen.

8. (a) To test the validity of the claim or assumption about the population parameter, a sample is drawn from the population and analysed. The results of the analysis are used to decide whether the claim is true or not. The steps of general procedure for any hypothesis testing are summarized below:

Step 1: State the Null Hypothesis (H_0) and Alternative Hypothesis (H_1): The null hypothesis H_0 (read as H_0 sub-zero) represents the claim or statement made about the value or range of values of the population parameter. The capital letter H stands for hypothesis and the subscript 'zero' implies 'no difference' between sample statistic and the parameter value. Thus hypothesis testing requires that the null hypothesis be considered *true (status quo or no difference)* until it is proved false on the basis of results observed from the sample data. The null hypothesis is always expressed in the form of mathematical statement which includes the sign ($\leq, =, \geq$) making a claim regarding the specific value of the population parameter. That is:

$$H_0: \mu (\leq, =, \geq) \mu_0$$

where μ is population mean and μ_0 represents a hypothesized value of μ . Only one sign out of $\leq, =$ and \geq will appear at a time when stating the null hypothesis

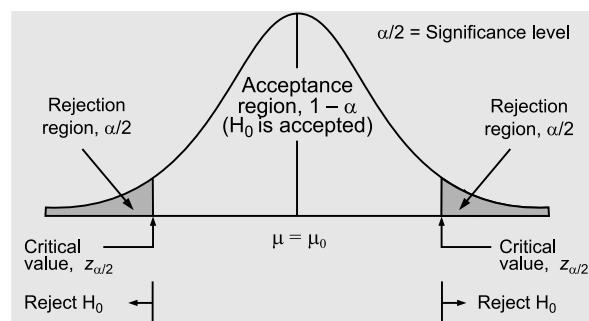
An **alternative hypothesis**, H_1 , is the counter claim (statement) made against the value of the particular population parameter. That is, an alternative hypothesis must be true when the null hypothesis is found to be false. In other words, the alternative hypothesis states that specific population parameter value is not equal to the value stated in the null hypothesis and is written as:

$$H_1: \mu \neq \mu_0$$

Consequently $H_1: \mu < \mu_0$ or $H_1: \mu > \mu_0$

Step 2: State the Level of Significance, α (alpha): The level of significance, usually denoted by α (alpha), is specified before the samples are drawn, so that the results obtained should not influence the choice of the decision-maker. It is specified in terms of the probability of null hypothesis H_0 being wrong. In other words, the level of significance defines the likelihood of rejecting a null hypothesis when it is true, i.e. it is *the risk a decision-maker takes of rejecting the null hypothesis when it is really true*. The guide provided by the statistical theory is that this probability must be ‘small’. Traditionally $\alpha = 0.05$ is selected for consumer research projects, $\alpha = 0.01$ for quality assurance and $\alpha = 0.10$ for political polling.

Step 3: Establish Critical or Rejection Region: The area under the sampling distribution curve of the test statistic is divided into two mutually exclusive regions (areas) as shown in figure. These regions are called the *acceptance region* and the *rejection (or critical) region*.



Step 4: Select the Suitable Test of Significance or Test Statistic: The tests of significance or test statistic are classified into two categories: *parametric and nonparametric tests*. Parametric tests are more powerful because their data are derived from interval and ratio measurements. Nonparametric tests are used to test hypotheses with nominal and ordinal data. Parametric techniques are the tests of choice provided certain assumptions are met.

Nonparametric tests have few assumptions and do not specify normally distributed populations or homogeneity of variance.

Selection of a test. For choosing a particular test of significance following three factors are considered:

- Whether the test involves one sample, two samples, or k samples?
- Whether two or more samples used are independent or related?
- Is the measurement scale nominal, ordinal, interval, or ratio?

One-sample tests are used for single sample and to test the hypothesis that it comes from a specified population.

The value of test statistic is calculated from the distribution of sample statistic by using the following formula

$$\text{Test statistic} = \frac{\text{Value of sample statistic} - \text{Value of hypothesized population parameter}}{\text{Standard error of the sample statistic}}$$

The choice of a probability distribution of a sample statistic is guided by the sample size n and the value of population standard deviation σ .

Step 5: Formulate a Decision Rule to Accept Null Hypothesis: Compare the calculated value of the test statistic with the critical value (also called *standard table value* of test statistic). The decision rules for null hypothesis are as follows:

- Accept H_0 if the test statistic value falls within the area of acceptance.
- Reject otherwise

In other words, if the calculated absolute value of a test statistic is less than or equal to its critical (or table) value, then accept the null hypothesis, otherwise reject it.

8. (b) Given $n = 200$

$$\begin{aligned} X &= \text{No. of faculty equipments} \\ &= 18 \end{aligned}$$

$$\begin{aligned} p &= \text{Sample proportion of faculty equipments} \\ &= \frac{X}{n} = \frac{18}{200} = 0.09 \end{aligned}$$

Null Hypothesis H_0 : $P = 0.95$ i.e. proportion of equipments confirmed to specification.

Alternative Hypothesis H_1 : $P \neq 0.95$ i.e. proportion of equipments not confirmed to specifications

Test Statistic:

$$\begin{aligned} Z &= \frac{p - P}{\sqrt{\frac{PQ}{n}}} \sim N(0, 1) & P &= 0.95 \\ & & Q &= 1 - 0.95 \\ & & &= 0.05 \\ &= \frac{0.09 - 0.95}{\sqrt{\frac{(0.95)(0.05)}{200}}} &= \frac{-0.86}{\sqrt{0.0002}} &= -60.99 \end{aligned}$$

$$|Z_{\text{cal}}| = 60.99$$

$$Z_{\text{tab}} = 1.96$$

Since $Z_{\text{cal}} (= 60.99) > Z_{\text{tab}} (= 1.96)$ we reject the null hypothesis at 5% level of significance and conclude that the proportion of equipments (95%) are not confirmed to specifications.

8. (c) **Null hypothesis H_0 :** The absence is uniformly distributed over the week

Days	Mon	Tue	Wed	Thu	Fri	Total
No. of absentees	59.8	59.8	59.8	59.8	59.8	299

$$\chi_{\text{cal}}^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$\frac{\text{Total}}{\text{No. of days}} = \frac{299}{5} = 59.8$$

O_i	E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
66	59.8	38.44	0.6428
56	59.8	14.44	0.2415
54	59.8	38.64	0.5625
48	59.8	139.24	2.3284
75	59.8	231.04	3.8635
			7.6387

$$\chi_{\text{cal}}^2 = 7.6387$$

$$\chi_{\text{tab}}^2 = \chi_{n-1}^2 = \chi_{5-1}^2 = \chi_4^2 = 9.49$$

Since $\chi_{\text{cal}}^2 (= 7.6387) < \chi_{\text{tab}}^2 (= 9.49)$ we accept the null hypothesis at 5% level of significance and hence conclude that the absence is uniformly distributed over the week.