

Visit us at www.himpub.com/manual/pcp080
for FREE Online Instructor and Student Manual

Quantitative Techniques for Business Managers

R.K. Bharadwaj
Anuradha R. Chetiya
Kakali Majumdar

Himalaya Publishing House

ISO 9001:2008 CERTIFIED

**QUANTITATIVE TECHNIQUES
FOR
BUSINESS MANAGERS**

QUANTITATIVE TECHNIQUES FOR BUSINESS MANAGERS

R. K. Bharadwaj

Director
Institute of Management Studies
Ghaziabad

Anuradha Rajkonwar Chetiya

Lecturer
Ramjas College
University of Delhi

Kakali Majumdar

Assistant Professor
College of Management
Shri Mata Vaishno Devi University



Himalaya Publishing House

ISO 9001:2008 CERTIFIED

© Authors

No part of this publication should be reproduced, stored in a retrieval system, or transmitted in any form or any means, electronic, mechanical, photocopying, recording and/or otherwise without the prior written permission of the publisher.

First Edition : 2009
Edition : 2011, 2014, 2016
Edition : 2017

Published by : Mrs. Meena Pandey for **Himalaya Publishing House Pvt. Ltd.**,
"Ramdoot", Dr. Bhalerao Marg, Girgaon, **Mumbai - 400 004.**
Phone: 022-23860170/23863863, Fax: 022-23877178
E-mail: himpub@vsnl.com; Website: www.himpub.com

Branch Offices :

New Delhi : "Pooja Apartments", 4-B, Murari Lal Street, Ansari Road, Darya Ganj,
New Delhi - 110 002. Phone: 011-23270392, 23278631; Fax: 011-23256286

Nagpur : Kundanlal Chandak Industrial Estate, Ghat Road, Nagpur - 440 018.
Phone: 0712-2738731, 3296733; Telefax: 0712-2721216

Bengaluru : Plot No. 91-33, 2nd Main Road Seshadripuram, Behind Nataraja Theatre,
Bengaluru - 560 020. Phone: 08041138821, 09379847017, 09379847005

Hyderabad : No. 3-4-184, Lingampally, Besides Raghavendra Swamy Matham, Kachiguda,
Hyderabad - 500 027. Phone: 040-27560041, 27550139

Chennai : New No. 48/2, Old No. 28/2, Ground Floor, Sarangapani Street,
T. Nagar, Chennai - 600 012. Mobile: 09380460419

Pune : First Floor, "Laksha" Apartment, No. 527, Mehunpura, Shaniwarpeth
(Near Prabhat Theatre), Pune - 411 030. Phone: 020-24496323/24496333;
Mobile: 09370579333

Lucknow : House No. 731, Shekhupura Colony, Near B.D. Convent School, Aliganj,
Lucknow - 226 022. Phone: 0522-4012353; Mobile: 09307501549

Ahmedabad : 114, "SHAIL", 1st Floor, Opp. Madhu Sudan House, C.G. Road, Navrang Pura,
Ahmedabad - 380 009. Phone: 079-26560126; Mobile: 09377088847

Ernakulam : 39/176 (New No.: 60/251) 1st Floor, Karikkamuri Road, Ernakulam,
Kochi - 682 011. Phone: 0484-2378012, 2378016; Mobile: 09387122121

Bhubaneswar : 5 Station Square, Bhubaneswar - 751 001 (Odisha).
Phone: 0674-2532129, Mobile: 09338746007

Kolkata : 108/4, Beliaghata Main Road, Near ID Hospital, Opp. SBI Bank,
Kolkata - 700 010, Phone: 033-32449649, Mobile: 07439040301

Printed at : Geetanjali Press Pvt. Ltd., Nagpur, On behalf of HPH.

PREFACE

This book aims to cover the applications of quantitative techniques in business management. An effort has been made to provide an extensive coverage of the latest quantitative methods that can aid a manager in decision making. The book is primarily aimed as a textbook of the paper on quantitative techniques of the MBA, PGDBM, B.Com (h), BBA, MIB courses of leading Indian Universities and Management Institutions.

This highlight of this book is its simply style of presenting perceived complex material. The communication and writing style is extremely student friendly with over 300 business applications used to explain different statistical concepts. We have also introduced relevant caselets and more examples as exercises for the students at the end of each chapter, to provide reinforcement to the applications of statistical techniques in business. Another factor, which differentiates this book, is the inclusion of an additional chapter on basic mathematical concepts, which have applications in management.

A positive feature of the book that will give a definite edge to students is that there is more focus on applications to explain different and difficult statistical concepts. Since the students of MBA and PGDBM come from diverse streams like the social sciences, care has been taken to present concepts in a simple manner, which they can easily comprehend. The coverage is extensive and the examples and caselets will help students relate to the examples in the Indian business context.

The Excel guide at the end of almost all the chapters is a unique feature of this book and gives it a distinct competitive advantage. This has been included to make students familiar with the use of statistical techniques on the computer using MS Excel.

Any comments, observations and/or suggestions to improve the book are most welcome. The authors can be reached at the following email ids: rkbharadwat1@hotmail.com, anuradha_rc@hotmail.com and kakali_m@yahoo.com.

Authors

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to **Mr. Pramod Aggarwal**, Chairman, IMS Group of Institutions for his unstinting support and encouragement.

A special note of thanks goes out to:

Gurpreet Arora and Bhandu for typing and editing the manuscript.

Pramod Sharma for facilitating the editing work between the publisher and the authors.

Manish Chetiya for chipping in and contributing to organize the Excel Guides.

A special word of thanks also goes out to:

Mr. Anuj Pandey, Mr. S. K. Joshi, Mr V. S. Rawat and the entire production team at the House for making this book become a reality.

CONTENTS

1. INTRODUCTION	1 - 14
1.1 An overview of Quantitative Techniques	
1.2 Quantitative Techniques and Decision Making in Business	
1.3 Quantitative Techniques and Computer Software	
1.4 Summary of the Chapters and Additional Features of this Book	
2. BASIC MATHEMATICS IN BUSINESS	15 - 52
2.1 Business Functions	
2.1.1 Linear and Quadratic Functions	
2.1.2 Cost, Revenue and Profit functions	
2.1.3 Break Even Point (BEP)	
2.2 Matrix Theory: Applications	
2.2.1 Definition and some types of Matrices	
2.2.2 Matrix Operations	
2.2.3 Applications of Matrices in Business Problems	
2.3 Differential Calculus: Applications	
2.3.1 Overview of Differential Calculus	
2.3.2 Applications of Derivative	
2.3.2.1 Price Elasticity of Demand	
2.3.2.2 Price Elasticity of Supply	
2.3.2.3 Average Cost and Marginal Cost	
2.3.2.4 Maxima & Minima	
2.4 Exercises	
3. COLLECTION AND PRESENTATION OF DATA	53 - 96
3.1 What is Data	
3.2 Types of Data and its Sources	
3.3 Collection of Primary Data	
3.3.1 Designing a Questionnaire	
3.4 Presentation of Data	
3.4.1 Ordered Array	
3.4.2 Pictorial Presentation of Data	
3.4.2.1 Diagrammatic Representation of Data	
Bar Diagram	
Pie Diagram	
Pictogram	
Population Pyramid	
Flowchart	

- 3.4.2.2 Graphical Representation of Data
 - Graphs of time series or line graphs
 - Area Graph
 - Scatter graph
 - Graphs of frequency distribution
- 3.4.3 Frequency Distribution
- 3.4.4 Relative Frequency Distribution
- 3.4.5 Cumulative Frequency
- 3.4.6 Graphical Representation of Frequency Distribution
- 3.4.7 Stem and Leaf Display
- 3.5 Caselet
- 3.6 Excel Guide
- 3.7 Exercises

4. MEASURES OF CENTRAL TENDENCY AND VARIATION

97 - 169

- 4.1 Introduction
 - 4.1.1 Concept of Central Tendency
- 4.2 Measures of Central Tendency
 - 4.2.1 Arithmetic Mean
 - 4.2.1.1 Properties of Arithmetic Mean
 - 4.2.1.2 Calculation of A.M from Ungrouped Frequency Distribution
 - 4.2.1.3 Calculation of A.M from Grouped Frequency Distribution
 - 4.2.1.4 Calculation of Weighted Mean
 - 4.2.1.5 Correction of Incorrect Observation
 - 4.2.1.6 Mean of Composite Group
 - 4.2.2 Harmonic Mean
 - 4.2.3 Geometric Mean
 - 4.2.3.1 Calculation of Geometric Mean Using Logarithms
 - 4.2.3.2 Combined Geometric Mean
 - 4.2.3.3 Weighted Geometric Mean
 - 4.2.4 Median
 - 4.2.4.1 Computation of Median
 - 4.2.4.2 Quartiles, Deciles and Percentiles
 - 4.2.4.3 Locating Quartile, Deciles and Percentiles Graphically
 - 4.2.5 Mode
 - 4.2.5.1 Computation of Mode
 - 4.2.6 Comparing the Mean, the Median and the Mode
- 4.3 Concept of Variation
- 4.4 Absolute Measures of Variation
 - 4.4.1 Range
 - 4.4.2 Quartile Deviation
 - 4.4.3 Mean Deviation (MD) or Mean Absolute Deviation (MAD)
 - 4.4.4 Standard Deviation (SD)

- 4.4.4.1 Important Properties of Dispersion
- 4.4.4.2 Calculation of Standard Deviation
- 4.4.4.3 Combined Standard Deviation
- 4.5 Relative Measures of Variation
 - 4.5.1 Coefficient of Variation
 - 4.5.2 Coefficient of Quartile Deviation
 - 4.5.3 Coefficient of Mean Deviation
- 4.6 Skewness
- 4.7 Kurtosis
- 4.8 Caselets
- 4.9 Excel Guide
- 4.10 Exercises

5. PROBABILITY AND PROBABILITY DISTRIBUTIONS

170 - 270

- 5.1 Introduction
- 5.2 Set Theory
 - 5.2.1 Definition of a Set
 - 5.2.2 Types of Sets
 - 5.2.3 Pictorial Representation of Set Theory - Venn Diagrams
 - 5.2.4 Properties of Set Operation
- 5.3 Counting Rules
 - 5.3.1 Permutations
 - 5.3.2 Combinations
- 5.4 Some Important Terms in Probability
 - 5.4.1 A Random Experiment
 - 5.4.2 Sample Space
 - 5.4.3 Trial
 - 5.4.4 Event
 - 5.4.4.1 Rules of Event Operations
 - 5.4.4.2 Types of Events
- 5.5 Various Definitions of Probability
 - 5.5.1 The Theoretical Definition of Probability
 - 5.5.2 The Classical Theory of Probability
 - 5.5.3 Relative Frequency or Empirical Approach
 - 5.5.4 Axiomatic Approach
 - 5.5.5 Subjective Approach
- 5.6 Laws and Theorems of Probability
 - 5.6.1 Additive law
 - 5.6.2 Conditional Probability and Multiplication Law of Probability
 - 5.6.3 Theory of Independence
 - 5.6.4 Pair-wise and Mutual Independence
 - 5.6.5 The Theorem of Total Probability and the Baye's Theorem

- 5.7 Probability Distribution of a Random Variable
 - 5.7.1 Random Variables
 - 5.7.2 Discrete Random Variable
 - 5.7.3 Continuous Random Variable
 - 5.7.4 Probability Distribution of a Random Variable
- 5.8 Discrete Probability Distributions
 - 5.8.1 Expected Value and Variance of a Discrete Probability Distribution
 - 5.8.2 Binomial Distribution
 - 5.8.2.1 Mean of Binomial Distribution
 - 5.8.2.2 Variance of Binomial Distribution
 - 5.8.2.3 Mode of the Binomial Distribution
 - 5.8.2.4 Fitting of Binomial Distribution
 - 5.8.3 Poisson Distribution
 - 5.8.3.1 Mean of Poisson Distribution
 - 5.8.3.2 Variance of Poisson Distribution
 - 5.8.3.3 Mode of the Poisson Distribution
 - 5.8.3.4 Poisson Approximation to Binomial
 - 5.8.3.5 An Application of the Poisson Distribution
 - 5.8.3.6 Fitting of Poisson Distribution
- 5.9 Continuous Probability Distribution
 - 5.9.1 Normal Distribution
 - 5.9.2 Characteristics of Normal Distribution
 - 5.9.3 Normal as an Approximation to Binomial Distribution
- 5.10 Caselets
- 5.11 Excel Guide
- 5.12 Exercises

6. SAMPLING AND SAMPLING DISTRIBUTIONS

271 - 315

- 6.1 Introduction
- 6.2 Parameter and Statistic
- 6.3 Sampling: Meaning, Steps and Types of Sampling
 - 6.3.1 Probability Sampling Methods
 - 6.3.1.1 Simple Random Sampling
 - 6.3.1.2 Stratified Random Sampling
 - 6.3.1.3 Systematic Sampling
 - 6.3.1.4 Cluster Sampling
 - 6.3.1.5 Multistage Sampling
 - 6.3.2 Non - Probability Sampling Methods
 - 6.3.2.1 Judgement Sampling
 - 6.3.2.2 Convenience Sampling
 - 6.3.2.3 Quota Sampling
 - 6.3.3 Sampling and Non-Sampling Errors

- 6.4 Sampling Distributions
 - 6.4.1 The Central Limit Theorem
 - 6.4.2 Sampling Distribution of the Mean
 - 6.4.3 Sampling Distribution of the Proportion
 - 6.4.4 Student's t-Statistic and its Distribution
 - 6.4.5 The Chi-Square Statistic and its Distribution
 - 6.4.6 The F-Statistic and its Distribution
- 6.5 Exercises

7. THEORY OF ESTIMATION AND TESTING OF HYPOTHESIS

316 - 391

- 7.1 Introduction
- 7.2 Estimation
 - 7.2.1 Point Estimation
 - 7.2.1.1 Point Estimator of Population Mean
 - 7.2.1.2 Point Estimator of Population Proportion
 - 7.2.1.3 Point Estimator of Population Variance
 - 7.2.2 Interval Estimation
 - 7.2.2.1 Interval Estimator of Population Mean
 - 7.2.2.2 Interval Estimator of Difference of Two Means
 - 7.2.2.3 Interval Estimator of Single Population Proportion
 - 7.2.2.4 Interval Estimator of Difference of Two Proportions
 - 7.2.2.5 Determination of Sample Size
- 7.3 Testing of Hypothesis
 - 7.3.1 Null and Alternative Hypothesis
 - 7.3.2 Type I Error Type II Error
 - 7.3.3 One-Tailed Test Two-Tailed Test
 - 7.3.4 One Sample Tests
 - 7.3.4.1 One Sample Z Test for Mean
 - 7.3.4.2 One Sample t Test for Mean
 - 7.3.4.3 One Sample Z Test for Proportion
 - 7.3.5 Two Sample Tests
 - 7.3.5.1 Two Sample Z Test for Difference of Two Means
 - 7.3.5.2 Two Sample t Test for Difference of Two Means
 - 7.3.5.3 Paired t Test (for correlated samples)
 - 7.3.5.4 Two Sample Z Test for Difference of Two Proportions
- 7.4 Caselets
- 7.5 Excel Guide
- 7.6 Exercises

- 8.1 Correlation Analysis
 - 8.1.1 Graphical Representation of Correlation
 - 8.1.2 Covariance
 - 8.1.3 Correlation Coefficient
 - 8.1.3.1 Karl Pearson's Correlation Coefficient
 - 8.1.3.2 Properties of Correlation Coefficient
 - 8.1.3.3 Standard Error and Probable Error of Correlation Coefficient
 - 8.1.4 Coefficient of Determination
 - 8.1.5 Rank Correlation
 - 8.1.6 Partial Correlation
 - 8.1.7 Multiple Correlation
 - 8.1.8 Testing the Significance of Correlation Coefficient
 - 8.1.9 Testing the Significance of Partial Correlation Coefficient
 - 8.1.10 Testing the Significance of Multiple Correlation Coefficient
- 8.2 Regression Analysis
 - 8.2.1 Simple Linear Regression
 - 8.2.1.1 Regression Equation and Regression Coefficients
 - 8.2.1.2 Properties of Regression Coefficients
 - 8.2.1.3 Least Square Method and Regression Equation
 - 8.2.1.4 Explained and Unexplained Variation
 - 8.2.1.5 Standard Error of Estimate
 - 8.2.1.6 Testing the Significance of Regression Coefficients
 - 8.2.2 Multiple Regression
 - 8.2.2.1 Multiple Regression with two Independent Variables
 - 8.2.2.2 Regression with Dummy Variable
- 8.3 Caselet
- 8.4 Excel Guide
- 8.5 Exercises

- 9.1 Introduction
- 9.2 Goal of Time Series Analysis
- 9.3 Components of Time Series
 - 9.3.1 Secular Trend
 - 9.3.2 Seasonal Component
 - 9.3.3 Cyclical Component
 - 9.3.4 Random/Irregular Component
 - 9.3.5 Models of Time Series
- 9.4 Measurement of Secular Trend
 - 9.4.1 Free Hand Curve Fitting
 - 9.4.2 Semi Average Method
 - 9.4.3 Moving Average Method

- 9.4.4 Fitting of a Straight Line/Trend Line
- 9.4.5 Fitting of Exponential Trend
- 9.4.6 Fitting of Second Degree Polynomial Equation
- 9.5 Measurement of Seasonal Component
 - 9.5.1 Method of Simple Average
 - 9.5.2 Ratio-to-Trend Method/Percentage-to-Trend-Method
 - 9.5.3 Ratio-to-Moving Average Method
- 9.6 Measurement of Cyclical Component
 - 9.6.1 Residual Method
- 9.7 Measurement of Irregular Component
- 9.8 Business Forecasting: An application of Time Series Analysis
 - 9.8.1 The Exponential Smoothing Method
 - 9.8.2 Trend adjusted for Seasonal Index Method
- 9.9 Caselets
- 9.10 Excel Guide
- 9.11 Exercises

10. CHI-SQUARE AND ANALYSIS OF VARIANCE

522 - 587

- 10.1 Introduction
- 10.2 The Chi – Square Statistic
 - 10.2.1 Chi-Square Test for Equality of Population Proportions or Chi – Square test for Homogeneity
 - 10.2.2 Chi- Square Test for Independence of Two Attributes
 - 10.2.3 Yates Correction for Continuity
 - 10.2.4 Chi- Square Test of Goodness of Fit
- 10.3 One way ANOVA
- 10.4 Assumptions of ANOVA
- 10.5 Simple Steps for ANOVA Calculations
- 10.6 Caselets
- 10.7 Excel Guide
- 10.8 Exercises

11. NON-PARAMETRIC TESTS

588 - 624

- 11.1 Introduction
- 11.2 One – Sample Runs Test
 - 11.2.1 Runs Test for Small Samples
 - 11.2.2 Runs Test for Large Samples
- 11.3 The Sign Test for Paired Observations
- 11.4 Rank Sum Tests
 - 11.4.1 Mann – Whitney U-test
 - 11.4.2 The Kruskal – Wallis H-Test
- 11.5 The Kolmogorov – Smirnov Goodness-of-fit Test
- 11.6 Exercises

1

Introduction



Structure →

- 1.1 An overview of Quantitative Techniques
- 1.2 Quantitative Techniques and Decision Making in Business
- 1.3 Quantitative Techniques and Computer Software
- 1.4 Summary of the Chapters and Additional Features of this Book

1.1 AN OVERVIEW OF QUANTITATIVE TECHNIQUES

Quantitative Techniques have long been a part of any decision making in business. In an age of information overload, companies today have access to an enormous amount of information whether qualitative or quantitative. However, information alone is not enough. What is required today is useful information. And information can only be useful when they are analyzed scientifically with the right set of tools to yield optimal useful information, that will help a company do better business.

Most quantitative and certain qualitative information relates to various kinds of data and quantitative techniques are very useful to analyze such data and make important conclusions and decisions. While there does not exist any universal definition of quantitative techniques, in general these techniques primarily include statistical techniques and the operations research techniques. In this book we discuss the statistical techniques of analyzing data with a focus on the practical applications of the techniques.

1.2 QUANTITATIVE TECHNIQUES AND DECISION MAKING IN BUSINESS

Decision making – one of the most crucial part of management is greatly supported by quantitative techniques. As we know, management of a business organization consists of managing and integrating various departments like Human Resources, Marketing, Production, Finance, Quality Control and so on, each executing different functions. Quantitative Techniques play a significant role in each of these areas, perhaps more in certain areas than in others and thus ultimately influences the decision making process of a business organization.

In this section, we discuss briefly situations and examples in some of the branches mentioned above, where quantitative methods play an important role in taking business decisions in the face of uncertainty.

Marketing and Quantitative Techniques

Marketing managers heavily rely on marketing research techniques to make decisions. Often, marketing managers have to find suitable answers to questions like,

- What is the best target market for certain products?
- What are the consumer's expectations from a product/service?
- How much is a consumer willing to pay for a certain product?
- What would be the most effective method of advertising for a new product?

and so on.

And most of these questions can be scientifically answered by using the marketing research process that, according to **Boyd, Westfall and Stasch (2004)**, consists of seven steps viz.

- (i) Specifying research objectives
- (ii) Preparing a list of the needed information
- (iii) Designing the data collection project
- (iv) Selecting a sample type
- (v) Determining sample size

- (vi) Organizing and carrying out the fieldwork
- (vii) Analyzing the collected data and reporting the findings

Statistical techniques are used extensively in the marketing research process, right from designing the data collection project to analyzing the collected data. Infact most books on marketing research devote entire chapters in describing these techniques and their applications in typical marketing situations. Apart from the basic statistical methods like averages and measures of dispersion that are described in detail in chapter 3 of this book, some of the other commonly used techniques are sampling methods, estimation and testing of hypothesis and correlation and regression analysis.

An example: Factor Analysis

Market research companies often need to identify the most important factors that characterize a product or service. For example, automobile manufacturers are often interested to know what characteristics a potential customer desires in an automobile. A statistical technique called factor analysis is often used for this purpose. In this method, the researcher prepares a long list of statements, maybe 100 or more, related to various attributes of the car. Customers are then asked to respond to each statement the extent to which they agree or disagree on a suitable scale. Factor analysis is then used to analyze this data and thus identify the major characteristics desired by the customers.

Production Management

This is one area of management in which statistical techniques are often used. Manufacturing, Aggregate Planning, Inventory Control, Work Scheduling, Job Sequencing and Maintenance are some of the areas in Production Management where these techniques are used. Apart from decisions like what to produce and how much to produce, quality control is one area where the use of these techniques become inevitable. In fact, a lot of theoretical development of these techniques has taken place from applications in quality management. The basic purpose of quality control tools is to ensure that products and services conform to specifications. Statistical tools like histograms, Pareto charts, control charts, statistical estimation and testing, Design of Experiments are all commonly used by engineers in production departments. Statistical Quality Control itself is a vast area covering both product quality and process quality through the use of tools like control charts, acceptance sapling, OC curves, ASN curves, process capability analysis and so on.

An example: Controlling a Production Process

Consider a company manufacturing gears to be used in automobile engines. The company supplies the gears to leading automobile manufacturers. Most manufacturers have stringent quality control measures in place and require the gears to conform to specific standards to ensure almost zero defects. As such, the company in turn needs to continuously ensure that the production process is in control and monitor the number of defectives. The factory runs for 20 hours and shuts down for maintenance everyday for 4 hours. Every hour, a certain number of gears are inspected from each production lot and the number of defectives noted down. Statistical process control techniques are then used to analyze the data.

Finance and Quantitative Techniques

Financial Management deals with the efficient use of economic resources namely capital funds. There are different quantitative techniques applied in Finance extensively. Budgetary Control, Portfolio Management, Security Analysis, Ratio Analysis, Risk and Return Analysis etc all depends on various tools of quantitative techniques like average, standard deviation, correlation regression, index numbers etc.

An Example: Risk and Return Analysis

Risk is present in every decision making process whether a production manager selects equipment or a marketing manager decides on an advertising campaign or a financial manager a portfolio of securities, all of them face uncertain cash flow. The objective in decision making is not to eliminate risk but to properly assess it. The subject matter of risk management is inherently quantitative. For example when we invest in a stock the return from it can take various possible values- 5%, 10% — etc. Here we depend on probability concept. The expected rate of return is calculated by taking weighted average of all the possible returns multiplied by their respective probabilities. Risk refers to the dispersion of a variable. It is commonly measured by the variance which is the sum of the deviations of actual return from the expected return.

Human Resource Management and Quantitative Techniques

Human resource is one of the important inputs of any business organization. The role of Human Resource Management (HRM) is to plan, develop, and administer policies and programmes designed to make expeditious use of an organization's human resources. The ultimate objective of the branch is the optimum utilization of the internal customer i.e. man power, by focusing on their recruitment, job evaluation, training, salary, safety, wellness, benefits, employee motivation etc. In most of the mentioned cases of HRM, the use of quantitative techniques is observed frequently.

An Example: Job Evaluation

Job evaluation is a practical technique, designed to enable trained and experienced staff to judge the size of one job relative to others based on their worth to the organization. It does not directly determine salary levels, but establishes the basis for an internal ranking of jobs. This method also enables the organization to compete in the market place for the best available talent and also allows the employees to compare the pay received with that received by employees of other organizations. Regression method is used in general to set a scientific job evaluation technique by the organizations. For example consider the following multiple regression model:

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_nX_n$$

where the dependent variable Y is the job value, X_i s are the responses of the different related questions like education level, year of experience etc.

Regression analysis of the above model helps the organization evaluate the value of the job from different angles. The standard error of the regression analysis will give the idea of the difference in predicted versus actual values of the dependent variable. Coefficient of determination will highlight the amount of variation in the dependent variable that is explained by the independent variables. The coefficients b_i s will indicate the weight of the questions considered in the analysis and so on.

1.3 QUANTITATIVE TECHNIQUES AND COMPUTER SOFTWARE

A number of statistical software packages are today available to make the use of the quantitative methods much simpler. Some of the commonly used packages in the industry are

- (i) Statistical Package for Social Sciences (SPSS)
- (ii) Statistical Analysis System (SAS)
- (iii) Minitab
- (iv) Matlab

(v) Microsoft Excel

(vi) Systat

It is possible to analyze all kinds of statistical databases with the help of these packages. Of course the interpretations of the outputs need to be taken special care of. A major advantage of these packages is that a person with limited knowledge of the theoretical mysteries behind the statistical tools can easily use them to analyze databases. The software does the entire analysis and produces the output and the user merely needs to interpret the results and effectively use them to make business decisions. Most of these software can also be used to conduct simulation studies and also write programmes.

We now present a snapshot of the above mentioned statistical packages:

(i) SPSS:

In 1968, Norman H. Nie, C. Hadlai (Tex) Hull and Dale H. Bent, developed SPSS based on the idea of using statistics to turn raw data into information essential to decision-making. Today, it is perhaps one of the most popular software for statistical data analysis.

An example: Regression Analysis using SPSS

To do a regression analysis on SPSS, the data is first typed in a SPSS data editor window.

Then from this window, go to *Analyze*, and then click on *Regression* and then *linear* as shown in Figure 1.1.

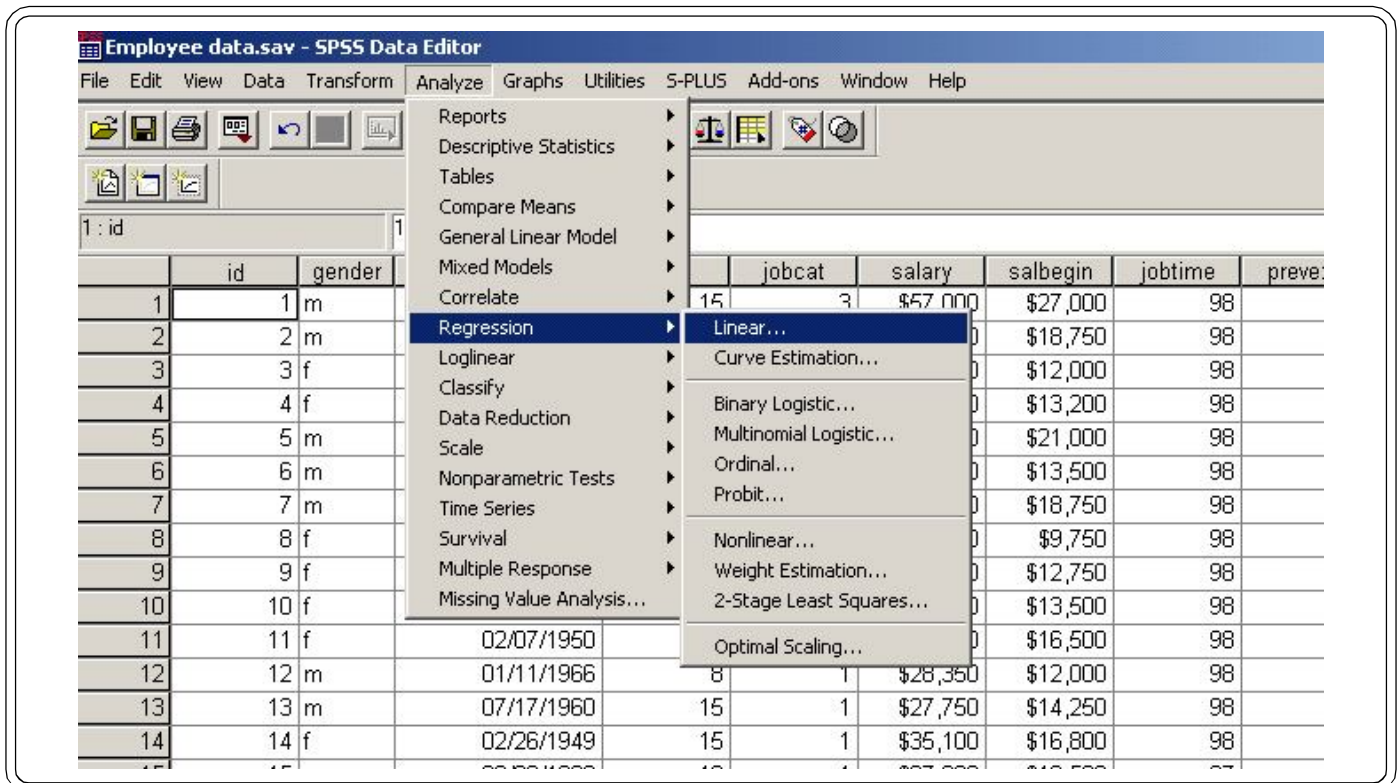


Figure 1.1

A SPSS Data Editor Window

Once the new window is open, select the dependent variable and the independent variable (Figure 1.2) and place them in their respective boxes on the right side of the window. Then click *Ok* to get the result sheet.

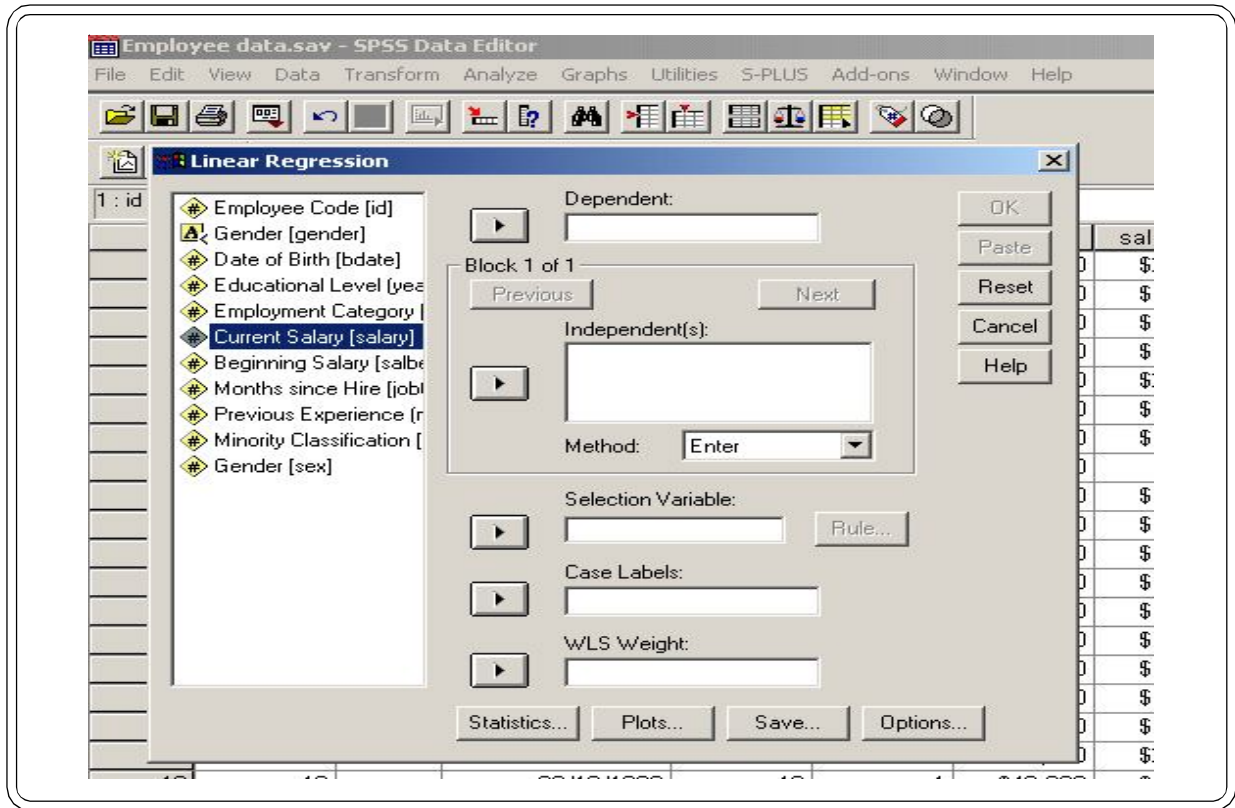


Figure 1.2

Linear Regression in SPSS

(ii) SAS

SAS is an integrated system of software products provided by SAS Institute that enables the programmer to perform many programs. SAS is driven by SAS programs that define a sequence of operations to be performed on data stored as tables. The main statistical functions of SAS are data entry, retrieval, management and mining, graphics, data warehousing, different statistical analysis etc. SAS is a programme based package and an example of a the SAS window is given in Figure 1.3.

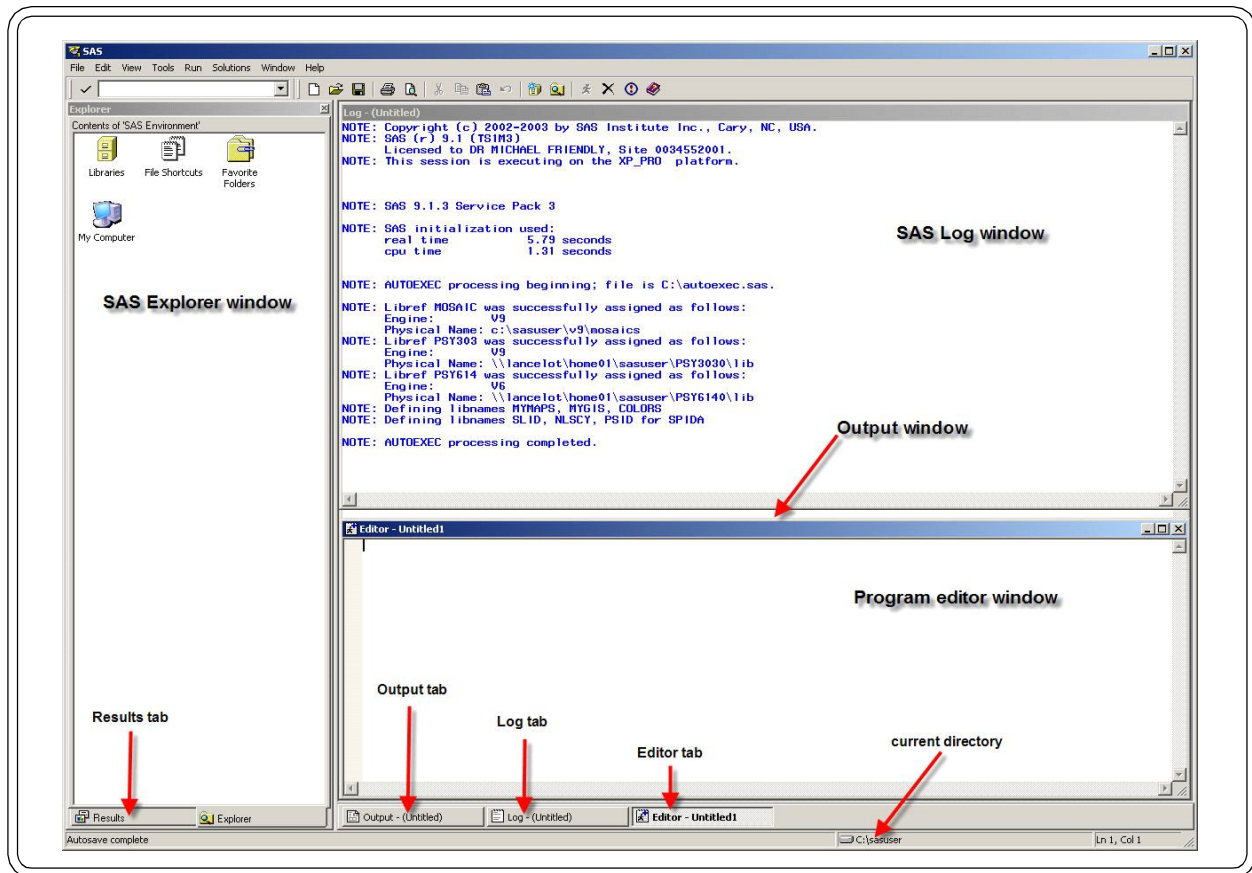


Figure 1.3

A SAS Window

(iii) Minitab

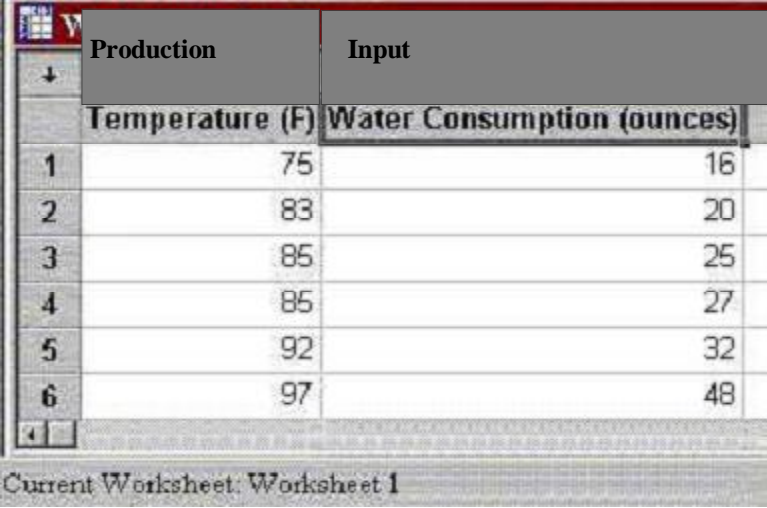
Minitab is also a popular statistical package with extensive industry usage. It was developed at the Pennsylvania State University by Barbara F. Ryan, Thomas A. Ryan, Jr. and Brian L. Joiner in 1972. Minitab began as a light version of OMNITAB, a statistical analysis program. Today, Minitab is often used in conjunction with the implementation statistics-based process improvement methods.

An example: **Calculation of Mean and Standard Deviation in Minitab**

To open Minitab the following steps are used:

First click the Start button in the bottom left hand corner of the screen.

Select Programs >Minitab for Windows>Minitab. Then Minitab will open. Enter the variables like



The screenshot shows a Minitab worksheet with two columns: 'Production' and 'Input'. The 'Production' column contains temperature values in degrees Fahrenheit, and the 'Input' column contains water consumption values in ounces. The data points are as follows:

	Production	Input
	Temperature (F)	Water Consumption (ounces)
1	75	16
2	83	20
3	85	25
4	85	27
5	92	32
6	97	48

Current Worksheet: Worksheet 1

Figure 1.4

Basic Statistics using Minitab

Select STAT > BASIC STATISTICS > DISPLAY DESCRIPTIVE STATISTICS.

In the Variables box, select Production, then click OK. The following output will be displayed.

Table 1.1
Descriptive Statistics: Production (Minitab Output)

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Production	7	88.00	85.00	88.00	8.47	3.20

(iv) MATLAB

MATLAB is a high-performance language for technical computation. Typically its uses include: Math and Computation, Algorithm development, Modeling, Simulation, Scientific and Engineering Graphics, Application development including Graphical User Interface building etc.

An example: Basic matrix operations using Matlab

Matlab can be extensively used for Matrix operations. We give here an elementary example of addition of two matrices.

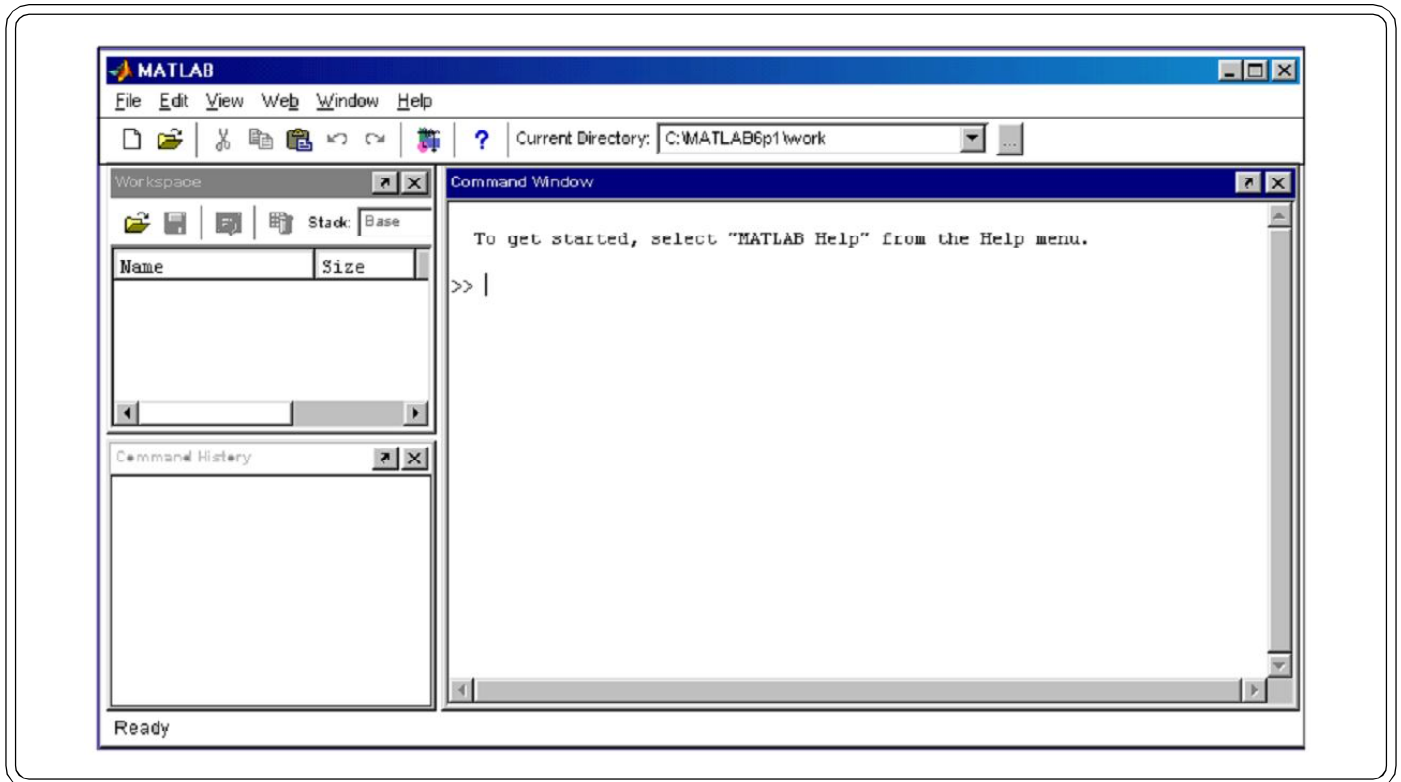


Figure 1.5

A MATLAB Window

The symbols for matrix addition, subtraction, multiplication and powers are +, -, * and ^
 If a is a square matrix then a^2 means $a*a$.

To enter the matrix $A = \begin{bmatrix} 1 & \dots & 2 \\ 3 & \dots & 4 \\ 5 & \dots & 6 \end{bmatrix}$ the following commands will be follows

```
>> b = [1 2; 3 4 ; 5 6 ]
```

Similar for $B = \begin{bmatrix} 3 & \dots & 4 \\ 0 & \dots & 1 \end{bmatrix}$

```
>> c = [3 4:0 1]
```

To add the matrices, the following command will be applied.

```
>> b + c
```

(v) Microsoft Excel

Microsoft office Excel or Microsoft Excel is a software program that allows the easy analysis and manipulation of data using tables and formulas. This book includes an EXCEL guide at the end of most of the chapters. The guide explains how the statistical techniques described can be easily calculated by using EXCEL spreadsheets.

An example: **Calculation of average using EXCEL**

Suppose from the EXCEL sheet we want to get the average of B3, C3 and D4 and the average will be displayed in column 5, i.e. E3.

Then click on cell E3 and write the following simple programme

= Average (B3:D3) and then click *Enter*.

The results will be displayed in column E (Figure1.)

An autofill option ensures that the rest of the cells are automatically filled up.

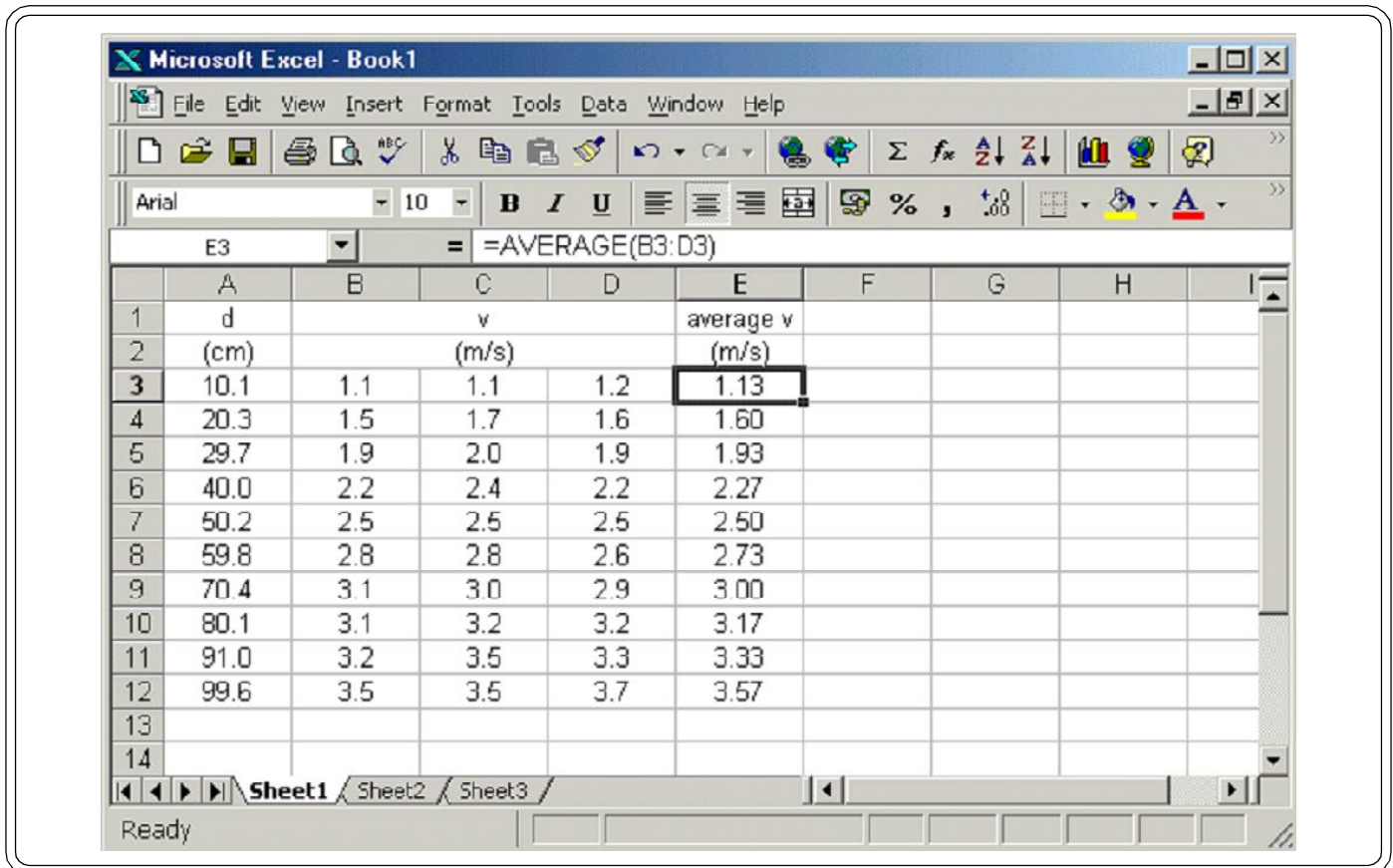


Figure 1.6

Calculation of average using EXCEL

(vi) SYSTAT

SYSTAT is a comprehensive, user friendly and highly integrated statistical software package most popularly used from microcomputers. SYSTAT includes the analysis of basic and advanced statistics.

An example: Calculation of Correlation using SYSTAT

From the Statistics menu select Correlations->Simple

Highlight the variables and click.

The SYSTAT window is shown in figure 1.7.

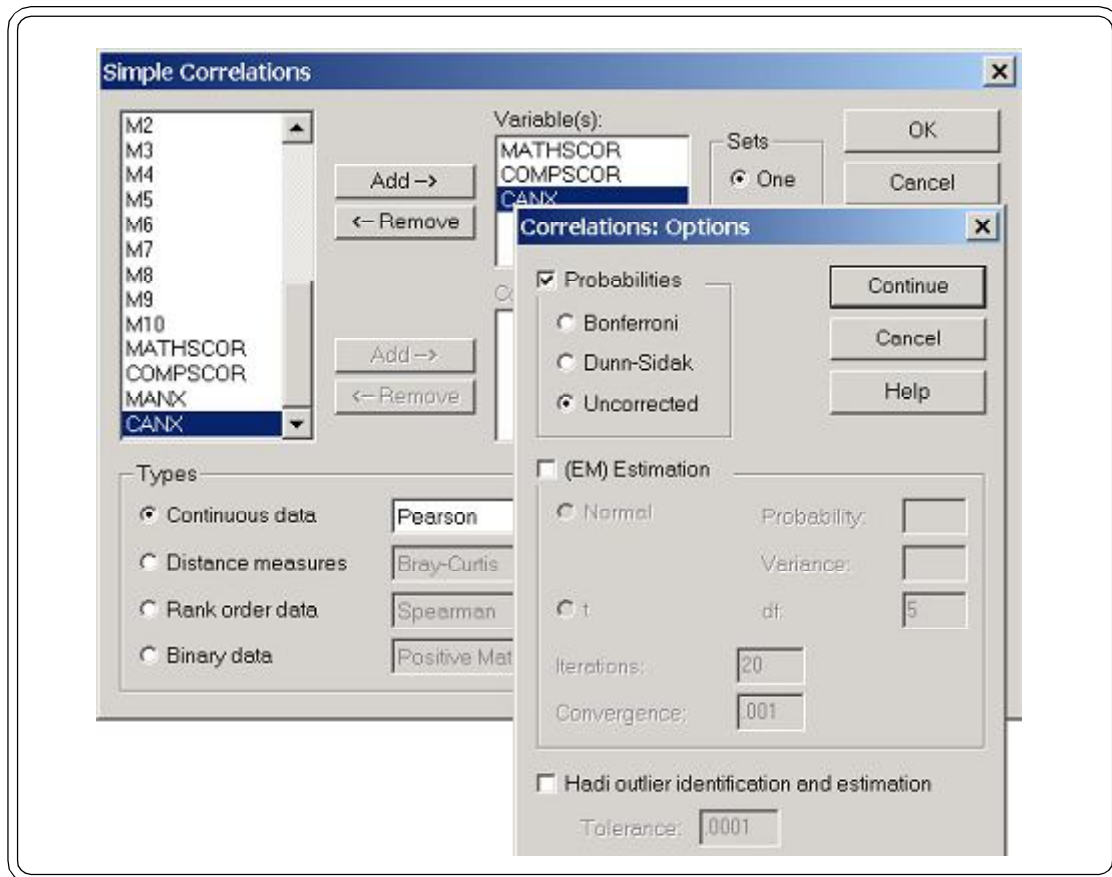


Figure 1.7

Calculating Correlation using SYSTAT

1.4 SUMMARY OF CHAPTERS AND ADDITIONAL FEATURES IN THIS BOOK

A brief synopsis of the chapters included in the book follows.

Chapter 2: Basic Mathematics in Business

The second chapter focuses on the use of mathematical concepts to simplify, understand and evaluate relationships between business variables. It includes definitions of simple functions like linear functions and quadratic functions and how they can be used to represent various cost functions, revenue functions and profit functions. It also includes the concept of break – even analysis along with numerous examples for each of these concepts. This chapter has a section of matrix theory and its applications in business situations. It includes definition of a matrix, its various types, rules of matrix operations before moving on to business examples simplified with matrices. For better understanding of the application numerous examples of different situations have been added along with clearly explained solutions. The last section of this chapter explains the concepts of differentiation, including in brief the basic concepts and rules of differentiation. In the application part it explains the linkage between differentiation and the average cost and marginal cost etc. and also how the concept of maxima and minima can be used to evaluate profits and losses of a company. This section also contains a lot of illustrations.

Chapter 3: Collection and Presentation of Data

Chapter 3 deals with data, the foundation of the entire area of statistics. Various types of data, different methods of data collection and their presentation both in pictorial and tabular form are explained here. The pictorial presentation is extensively researched and includes representation like box plot, pictogram, population pyramid, area graphs, stem and leaf diagram and flowcharts. A sufficient number of examples of each type have been solved.

The tabular presentation explains the basic forms of representing data in tabular form with lots of current examples. We have further tried to enhance this chapter by including an **Excel guide** and a contemporary caselet.

Chapter 4: Measures of Central Tendency and Dispersion

This chapter explains the different measures of central tendency and variation of data. The arithmetic mean, median, mode are defined and their computation have been explained with examples. In addition the geometric mean and the harmonic mean are defined with typical applications. The concept of variability and its importance in understanding the behavior of statistical data is the focus of defining and explaining the various measures of dispersion. A special emphasis is given on the variance and standard deviation that are central to the concept of data variability. In this chapter in addition to explaining the various measures with examples, the concept of quartiles, deciles and percentiles has been innovatively explained with the help of diagrams. We have included two caselets to help students understand the applicability of these measures in real life situations. Also, an Excel guide has been included explaining how these measures can be calculated by using **Microsoft Excel**.

Chapter 5: Probability and Probability Distributions

Both probability theory and probability distribution have been discussed in chapter 5. This chapter begins with a section on set theory, which is central to probability theory. The concept of probability, various definitions of probability and laws and theorems of probability are explained thereafter. Solved examples based on each of these areas are given in plenty. This is followed by a detailed discussion of the various important probability distributions like the Binomial distribution, the Poisson distribution and the ever-popular and widely applicable Normal distribution. Important properties of each distribution, their applicability in different situations are discussed with relevant examples. The **Excel guide** included in this chapter provides a step-by-step guide for calculation of probabilities of these distributions in Excel. This chapter also includes caselets related to real world problems.

Chapter 6: Sampling & Sampling Distributions

This chapter begins with a description of the different sampling methods. Each of these methods is highlighted with illustrations of typical situations when they can be applied. The sampling methods discussed here include simple random sampling, stratified random sampling, systematic sampling, cluster sampling and multistage sampling. Among the non-probability methods quota sampling, judgment sampling and convenience sampling are discussed. This is followed by a discussion on the much-celebrated Central Limit Theorem that forms the basis of sampling distributions, which in turn is the backbone of the entire inferential theory. Discussions on Sampling distribution of the mean, the t-statistic, the Chi-square Statistic and the F – Statistic follow.

Chapter 7: Estimation & Testing of Hypothesis

Since inferential statistic is a topic which students find difficult to comprehend we have tried our best to make this chapter student friendly. The chapter has been structured by dividing it into a number of sub-sections where the different types of estimation and testing has been discussed clearly with lots of suitable practical examples which students will be able to relate to. The first section focuses on the theory of estimation. In particular point estimators and interval estimators have been explained with simple and practical examples, which were well received by students when tried out in classrooms. The next section now tries to explain the importance of testing hypothesis made about these estimates. Basic terms and concepts related to hypothesis testing are explained in detail. The testing process has been structured in a series of steps and these steps are followed throughout while explaining the examples. This chapter has extensively covered all the tests for single and two samples with at least three typical examples for each test. Caselets, which is our modest attempt in this book, is also a part of this chapter. We hope it will make these concepts easy to understand in the real world context. An **Excel guide** also gives a step-by-step guide to test some of the hypothesis described in this chapter.

Chapter 8: Correlation and Regression

Till now we were dealing with univariate data. Discussion on bivariate and multivariate data starts from this chapter onwards beginning with a discussion on concepts of Correlation. Besides the Pearson's correlation, we have also covered coefficient of determination, rank correlation, partial and multiple correlations all with suitable examples. Testing for significance of correlation has been explained with illustrations. We hope these concepts will help the students visualize the extension of the concept of correlation from bivariate to multivariate data. The next section on regression analysis includes definition, concept of the two different regression lines, the least square method of calculation regression Standard error associated with the estimating equation and testing for significance of the regression coefficients. A section follows this on Multiple Regression. The topics covered in this section are multiple regression with two independent variables and multiple regression with dummy variables. The latter, in particular is used quite often by social scientists in their research. A caselet where some techniques discussed in this chapter can be applied has been given. The EXCEL guide given in this chapter demonstrates calculation of scatter diagram, correlation coefficient and simple linear regression using a EXCEL spreadsheet.

Chapter 9: Time Series and Forecasting

The discussion on bivariate data has been extended in the next chapter. The students are introduced to the concept of time series analysis and its application in business forecasting. Different concepts related to time series like components of time series and how to measure them can be found in this chapter. It includes a section each on linear and non-linear trend fitting respectively. Two small case studies have been given in this chapter for students to analyze using concepts explained in this chapter. Also, calculation of moving averages, trend values etc through an **Excel guide** have been given in concise steps.

Chapter 10: Chi – Square and Analysis of Variance

In chapter 10 our focus is to explain the concepts of chi square analysis and one-way ANOVA using as many examples as possible so that students can easily understand how and under what conditions to apply these tests. Of course the theoretical basis has been also explained in brief. All three applications of the chi square statistic viz. tests for homogeneity, tests for independence of attributes and tests for goodness of fit are explained using lots of illustrations. To further reinforce

their applicability two caselets have been carefully researched and included. The **Excel guide** provides the steps for calculation of Chi Square and ANOVA using an excel spreadsheet.

Chapter 11: Non Parametric Tests

Non-Parametric tests are often used in many business problems besides their applications in other areas primarily because they do not make any restrictive assumptions about the normality of the parent population and also because they are arguably computationally simpler. In this chapter we present a brief discussion on a few commonly used non parametric tests.

References

1. Harper W. Boyd, Jr, Ralph Westfall, Stanley F. Stasch (2004): Marketing Research: Text and Cases, Seventh Edition, Richard D. Irwin, Inc. Homewood, Illinois.



2

Basic Mathematics in Business



Structure

2.1 Business Functions

2.1.1 Linear and Quadratic Functions

2.1.2 Cost, Revenue and Profit functions

2.1.3 Break Even Point (BEP)

2.2 Matrix Theory: Applications

2.2.1 Definition and some types of Matrices

2.2.2 Matrix Operations

2.2.3 Applications of Matrices in Business Problems

2.3 Differential Calculus: Applications

2.3.1 Overview of Differential Calculus

2.3.2 Applications of Derivative

2.3.2.1 Price Elasticity of Demand

2.3.2.2 Price Elasticity of Supply

2.3.2.3 Average Cost and Marginal Cost

2.3.2.4 Maxima & Minima

2.4 Exercises

2.1 BUSINESS FUNCTIONS

In decision making problems, an important issue is the identification of relationships among the various decision variables. These relationships are usually in the form of an equation, a set of equations or inequalities. A Function can be used to describe certain quantitative relationships mathematically. This Chapter aims to explain some of the functions, which occur commonly in business like the linear function, cost function, revenue function and profit function. It also deals with cost minimization, revenue and profit maximization by using the mathematical concepts of derivatives. The section on matrices dwells on the wide-ranging applications of matrix theory in business situations. It explains how many business situations can be simplified by applying matrix theory.

2.1.1 Linear and Quadratic Functions

A linear function is of the form

$$y = a + bx$$

where x is the independent variable,
 y is the dependent variable and
 a & b are the constants to be determined.

This function is graphically depicted in figure 2.1.

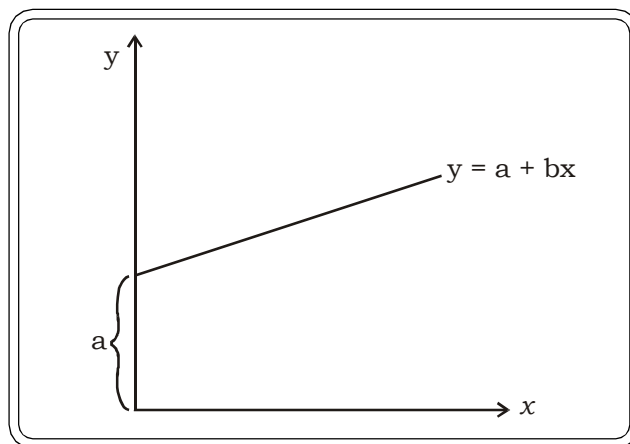


Figure 2.1

A Simple Linear Function

a here represents the y intercept and b the slope. b is also interpreted as the unit increase in y corresponding to a unit increase in x .

Let us try and understand the parameters a and b with the help of this simple example of a taxi meter. A taxi meter say, starts with an initial amount of Rs 30. For every additional km one travels, suppose one has to pay Rs 5. If we consider the number of kilometers travelled as x and the total fare as the variable y , we can express the relationship between the number of kilometers travelled and the amount we have to pay at the end of our journey.

This will be of the form $y = 30 + 5x$

This linear equation can now be used to find out the cost of traveling x kms by the taxi.

Example 2.1: In business, a linear function can be constructed from a situation like this. Suppose the fixed cost of producing x units of a commodity is Rs.10, 000 and thereafter the variable cost of producing each unit cost Rs.5. Then the total cost function is:

$$TC(x) = 10,000 + 5x$$

which is again a linear function.

Example 2.2: The life expectancy of a person in 1990 was 70 years. In 2000 it was 75 years. The life expectancy can now be expressed as a linear function of time. This equation can now be used to predict life expectancy 5 years from now or 10 years from now and so on. That is this line can now serve as an estimating equation.

If life expectancy is denoted by L , and the number of years by t , the estimating linear function now becomes:

$$L = a + bt$$

A method of estimating the parameters a and b is described in the chapter 8.

A Quadratic Function is of the form

$$y = a + bx + cx^2$$

where y is the dependent variable,

x is the independent variable

a , b , c are the parameter that can be estimated (methods of estimating a , b and c are discussed in chapters 8 and 9).

The graph of a quadratic equation is a parabola which opens upwards or downwards depending on the value of a . If $a > 0$, it opens upwards, the basic form being of a 'valley' (Figure 2.3) and if $a < 0$, it opens downwards the basic form that of a 'mountain' (Figure 2.2).

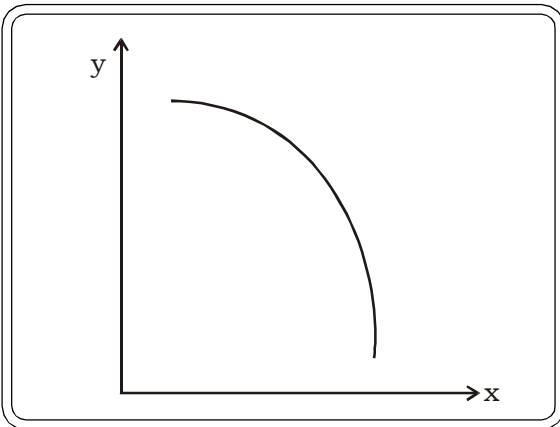


Fig. 2.2

Graph of $y = a + bx + cx^2$, when $a < 0$

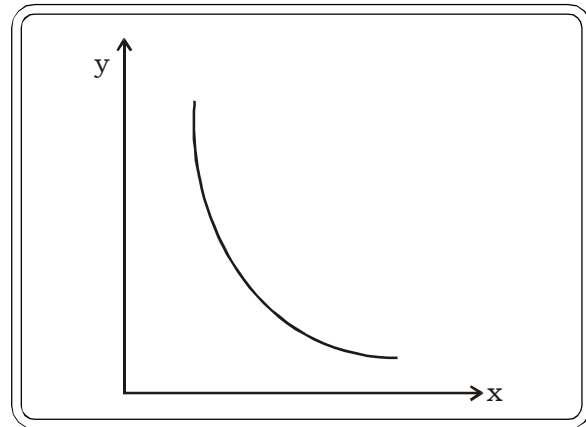


Fig. 2.3

Graph of $y = a + bx + cx^2$, when $a > 0$

The quadratic equation can also be used to express business functions like the profit function, revenue function and this is illustrated in the following example.

Example 2.3: A company's total profit (in Rs) over a particular period is given by $10x^2 - 6x - 2x^3$, where x represents the output volume. Then the company's profit per unit is given by:

$$y = \frac{10x^2 - 6x - 2x^3}{x} = 10x - 6 - 2x^2, \text{ which is a quadratic function.}$$

Linear and quadratic functions are commonly used in business in approximating revenue, cost and profit functions. We shall look at the applications in detail in the next section.

2.1.2 Cost, Revenue and Profit Functions

The basic quantitative models in business and economic applications are those involving the relationship between a variable related to volume such as production volume and cost revenue and profit. Linear and quadratic functions may be applied to establish relationships and through the use of these functions a manager can determine the projected cost, revenue and profit associated with a certain volume. All management departments like finance, production, sales etc can benefit in their areas of decision-making through the use of these cost, revenue and profit models. A closer study of these models now follows.

Cost related to Volume Function

The cost related to any production is a function of the amount produced or the volume of production. A linear function can be used to relate this. The cost component consists of two parts – fixed cost and a variable cost. The fixed cost is independent of the volume and the variable cost depends on the volume produced. For example if the fixed set up cost for production of a facility is Rs 40,000 and the variable cost of manufacturing is Rs 10 for each unit, then the cost volume model for production of x units would be the linear function:

$$\text{Cost (y)} = \text{Rs } 40,000 + \text{Rs } 10 x \text{ (x units)}$$

As defined before, b represents the unit increase in y corresponding to a unit increase in x . Thus here b is also the marginal cost i.e. the rate of change in the total cost with respect to volume. In this example Rs 10 is called the marginal cost – the cost of producing one additional unit.

Revenue and Volume Function

Also related to volume is another important parameter i.e. the revenue generated by selling a certain number of units. To model this relationship let us look at the following example.

If the product is to be sold at Rs.15 per unit the total revenue generated would be Revenue (y) = Rs.15 x , where x is the number of units sold.

This is again a linear model but we notice that the value of a is 0. This is because we do not start off with any fixed revenue. This is completely a function of the number of units that are sold.

In economics, marginal revenue is defined as the rate of change of total revenue with respect to sales volume. Thus in this equation the marginal revenue is Rs 15 which is the increase in total revenue resulting from the sale of an additional unit.

Profit and Volume Function

One of the primary drivers of business and the decision-making process is profit. The cost – volume and the revenue – volume functions can be now combined to give a profit model which will help a manager arrive at profit margins associated with a certain volume of sales. The profit function is the difference between the amount of revenue generated and the total cost. Mathematically this can be expressed as

$$\text{Profit} = \text{Revenue} - \text{Cost}$$

$$P(x) = R(x) - C(x) \quad \dots \text{(i)}$$

Combining the revenue and the cost functions in the example the profit function now becomes

$$\begin{aligned} P(x) &= 15x - (40,000 - 10x) \\ &= -\text{Rs.}40,000 + \text{Rs.}15x \end{aligned}$$

We now take a look at some detailed examples of the various models described above.

Example 2.4: A company manufactures x pens each day at a cost of Rs.10 each. The cost of manufacturing and selling the pens is Rs.5 plus a fixed daily overhead cost of Rs.800. Determine the profit function. What will be the profit if 500 pens are produced and sold every day?

Solution:

The revenue and volume function is:

$$R(x) = 10x$$

Total cost of manufacturing x pens a day is given by the cost volume model

$$C(x) = 800 + 5x.$$

The profit per day = Revenue - Cost = $R(x) - C(x) = 10x - (800 + 5x) = 5x - 800$

The profit when 500 pens are produced in a day = $2500 - 800 = \text{Rs. } 1700$

Example 2.5: A book publisher has calculated the production cost associated with each book as Rs. 50 and the fixed costs are Rs. 20, 000 if each book is sold for Rs. 75.

1. Determine the cost function.
2. Determine the revenue function.

Solution:

1. The cost function assuming x books are to be published is:

$$C(x) = \text{Rs. } 20,000 + \text{Rs. } 50x$$

2. The revenue function:

$$R(x) = \text{Rs. } 75x$$

Example 2.6: If the fixed costs of producing a product are Rs. 100 and the average variable costs is Rs. 5 and the selling price is Rs. 8, find

- (a) The cost of producing x units.
- (b) The revenue function.
- (c) The profit function.

Solution:

(a) the total cost of producing x units is given by the cost function $C(x) = 100 + 5x$

(b) the revenue function from selling x units is given by the revenue function $R(x) = 8x$.

(c) The profit from producing and selling x units is given by the profit function

$$\begin{aligned} P(x) &= R(x) - C(x) \\ &= 8x - (100 + 5x) \\ &= 3x - 100 \end{aligned}$$

Example 2.7: Sometimes the average variable cost may not be a constant, but may vary. This happens because suppliers often tend to give discounts for bulk orders. For example, this could imply that the more the quantity being ordered, the greater the discount.

Let the average variable cost be $2 - 0.1x$. Suppose the fixed cost remain Rs.100 and the unit-selling price is Rs. 8. Find the profit function.

Solution:

The cost function now becomes

$$C(x) = 100 + (2 - 0.01x)x = 100 + 2x - 0.01x^2$$

The revenue function $R(x) = 8x$

The profit function

$$\begin{aligned} P(x) &= 8x - (100 + 2x - 0.01x^2) \\ &= -100 + 0.05x^2 + 6x, \text{ which is a quadratic function} \end{aligned}$$

Example 2.8: If the market demand function of a product is $x = 160 + 8p$

where x – the quantity supplied

p – the market price

the unit cost of production is Rs.5. Determine at what price this product should be sold to achieve a profit of Rs.800.

Solution:

The total profit function

$$\begin{aligned} P(x) &= \text{Total revenue} - \text{Total cost} \\ &= px - 5x \end{aligned}$$

Expressed in terms of price this function becomes

$$\begin{aligned} P(p) &= p(160 + 8p) - 5(160 + 8p) \\ &= 160p + 8p^2 - 800 - 40p \end{aligned}$$

Setting the profit equal to 800, we get the value of p

$$\Rightarrow 800 = 8p^2 + 120p - 800$$

$$\Rightarrow 8p^2 + 120p - 1600 = 0$$

Solving this quadratic equation

$$p = 8.5, -23.5$$

Since the price cannot be negative we consider Rs 8.5 as the appropriate price at which to sell the product, to achieve a profit of Rs. 800.

Example 2.9: A man selling pots sells an average of 5 pots per day at a price of Rs 20 each. By increasing his sale to an average of 6 pots he could obtain Rs 18 each. Assume that the demand function is of the form

$$x = a + bp$$

where x – average number of pots sold per day

p – price of each pot & a and b are constants to be determined

His daily production costs are

$$c = \frac{1}{2}x^2 - \frac{1}{2}x + 54$$

(i) Determine the constants a & b and hence the demand function.

(ii) Determine the profit function.

Solution:

(i) The demand function is of the form

$$x = a + b p$$

To calculate a & b we use the given information related to price and sales to get the two equations

$$5 = a + b 20 \quad \dots (1)$$

$$6 = a + b 18 \quad \dots (2)$$

Solving these two equations

$$a = 15 \text{ and } b = -\frac{1}{2}$$

Thus the demand function is:

$$x = 15 - \frac{1}{2} p$$

(ii) The profit function

$$\begin{aligned} P(x) &= 2x(15 - x) - \left(\frac{1}{2}x^2 - \frac{1}{2}x + 54\right) \\ &= -1.5x^2 + 15.5x - 54 \end{aligned}$$

2.1.3 The Break Even Point (BEP)

The break-even point (BEP) simply defined is the point at which the total revenue is equal to the total cost of producing the product. The following example will help illustrate the concept of BEP.

The equation $R(x) - C(x)$ helps to determine any profit associated with a production volume.

Example 2.10: If 50 pens are produced the profit is say,

$$P(50) = 5(50) - 500 = -250$$

In other words, a loss of Rs.250

If sales are expected to be 50 units, the manager may end up not producing the product. However a demand of 400 units would yield a profit of

$$P(400) = 5(400) - 500 = 1500$$

This profit may justify the production of the pens.

Thus a volume of 50 units results in a loss and a volume of 400 units yields a profit of Rs.1500. The volume that results in total revenue equaling total cost is called the break-even point. Knowing a break even point is useful for the manager because he can quickly infer that a volume above it will result in profits and a volume below it will result in losses. For any new business venture it is important to be able to predict what must be the sales volume to reach the break-even point and subsequently start making profits. Thus a break-even point provides useful insights to a manager who must make a yes/no decision regarding manufacturing of any product.

In example 2.4 we now determine the BEP by setting $R(x) = C(x)$

$$10x = 800 + 5x \Rightarrow 5x = 800 \Rightarrow x = 160$$

This number tells us that the sales of the pens must be at least 160 before a profit can be expected. The graph depicting the BEP is shown in the following figure:

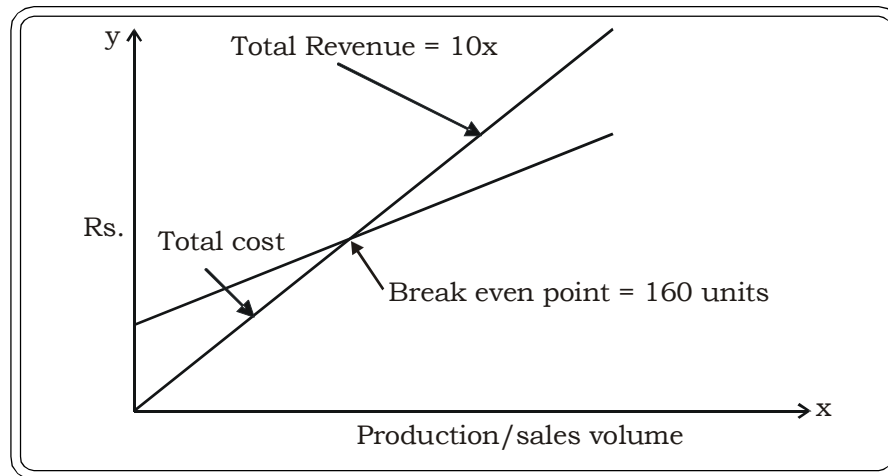


Figure 2.4

The Break Even Point

Example 2.11: The pricing policy of ABC company follows the demand equation $p = D(x)$, where $D(x)$ is the price per unit when x units are demanded. After studying the market trends, the company determines this function to be

$$D(x) = 3000 - 3x$$

The fixed costs are estimated to be Rs. 30,000 and in addition the company will have to pay Rs.400 for each unit produced. Determine the sales level at which ABC company can recover its costs.

Solution:

The Revenue Function: $R(x) = D(x) \cdot x = (3000 - 3x) \cdot x = 3000x - 3x^2$

The Cost Function: $C(x) = 30,000 + 400x$

For the break even point: $R(x) = C(x)$

$$3000x - 3x^2 = 30,000 + 400x$$

$$\Rightarrow 3x^2 + 400x - 27000 = 0$$

$$\Rightarrow x =$$

The company can recover costs by producing units of the product.

Example 2.12: A firm manufacturing compact discs (CD's) has the following costs.

Fixed costs: Rs.10,000 and variable costs: Rs5 per unit. The selling price of each CD is Rs.8. How many units of CD does the firm need to manufacture before it can start making profits?

Solution:

The cost function: $C(x) = 10,000 + 5x$

The revenue function: $R(x) = 8x$

For determining the BEP we set $C(x) = R(x)$

$$\Rightarrow 8x = 10,000 + 5x$$

$$\Rightarrow x = \frac{10000}{3}$$

$$\Rightarrow x \cong 3333$$

The company needs to sell at least 3333 CD's in order to break even.

Example 2.13: The price function of a profit making company is given by $p = 2000 - 3x$. The fixed cost for the company includes cost of rent & rates, depreciation, R & D, administration costs and is estimated at Rs.60,000. The variable cost is estimated to be Rs 500 for each unit that it produces. At what level of sales can this company recover its costs?

Solution:

The cost function: $C(x) = 60,000 + 500x$

The revenue function: $R(x) = px$
 $= (2000 - 3x)x$
 $= 2000x - 3x^2$

For determining the BEP we set $C(x) = R(x)$

$$\Rightarrow 2000x - 3x^2 = 60,000 + 500x$$

$$\Rightarrow 3x^2 - 1500x + 60,000 = 0$$

Solving this quadratic equation we get the values of x as

$$x = 44, 456.$$

Thus the company will recover their cost as long as the demand for the product is between the BEP's 44 and 456.

Example 2.14: The total daily cost of manufacturing computer tables is given by $Y = 2.5x + 300$.

- (i) If each table sells for Rs.4, find the break-even point.
- (ii) If the selling price is increased to Rs.6 per table what is the new break-even point?
- (iii) If it is known that at least 150 tables can be sold each day, what is the price to be charged to ensure that there is no loss.

Solution:

(i) The cost function: $Y = 300 + 2.5x = C(x)$ (say)

The revenue function: $R(x) = 4x$

For determining the BEP we set $C(x) = R(x)$

$$\Rightarrow 4x = 2.5x + 300$$

$$\Rightarrow 1.5x = 300$$

$$\Rightarrow x = 200, \text{ which is the BEP.}$$

(ii) If the selling price is increased to Rs.6 we now examine its impact on the break-even point

$$\Rightarrow 6x = 2.5x + 300$$

$$\Rightarrow 3.5x = 300$$

$$\Rightarrow x = 85$$

Thus, by increasing the sales price by Rs. 2, the manufacturer can break even by selling 85 tables.

(iii) To determine the price at which 150 tables should be sold to guarantee no loss:

$$675 = 150p$$

$$\Rightarrow p = \text{Rs.4.5}$$

Example 2.15: The fixed costs of producing a certain product are Rs.5000 per month and the variable cost is Rs.3.50 per unit. If the product sells for Rs.6 each, find the following

- (i) The break even point
- (ii) The number of units that must be produced and sold each month to obtain a profit of Rs.1000 per month
- (iii) The loss, when only 1500 units are produced and sold each month.

Solution:

(i) The cost function: $C(x) = 5000 + 3.5x$

The revenue function: $R(x) = 6x$

For determining the BEP we set $C(x) = R(x)$

$$\Rightarrow 6x = 3.5x + 5000$$

$$\Rightarrow 2.5x = 5000$$

$$\Rightarrow x = 2000$$

(ii) Profit function: $P = 6x - (5000 + 3.5x)$

$$\Rightarrow 1000 = 2.5x - 5000$$

$$\Rightarrow x = 2400$$

Thus 2400 units have to be produced and sold each month to generate a profit of Rs.1500.

(iii) When 1500 units are produced the loss amounts to

$$\Rightarrow P = 2.5(1500) - 5000 = -1250$$

Thus the loss is Rs.1250

2.2 MATRIX THEORY: APPLICATIONS

Matrices are very useful in practical business purposes whether it is finance, economics or linear programming to name just a few areas. Economists use matrices extensively in various fields like social accounting and in studying 'inter industry economics. Linear programming problems used extensively in operations research are rooted in matrix theory. In fact almost the entire discipline of Operations Research from LPP to game theory relies heavily on matrix operations and concepts for analysis. In this chapter we try to give a brief definition of certain important types of matrices, define their properties, look at an overview of the matrix operations and finally focus on some applications in business.

2.2.1 Definition and some types of Matrices

A matrix as all books would define is simply an arrangement of numbers in rows and columns and enclosed by a pair of brackets. However the opportunities that such an arrangement offers are infinite. A general definition of a matrix of order $m \times n$ is:

$$\begin{pmatrix} 2 & 5 & -1 \\ 3 & -6 & 4 \end{pmatrix}$$

This being a matrix of order 2×3 i.e. 2 rows and 3 columns

A general definition of a matrix of order $m \times n$ is $A = (a_{ij})_{m \times n}$, which can be written as.

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

Types of Matrices

The different types of commonly used matrices are

(i) Square Matrix

This is a matrix whose number of rows is equal to the number of columns. For a $m \times n$ matrix this implies $m = n$. Examples of square matrices of order 2 and 3 are

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \text{ and } \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

(ii) Row and Column Matrices

A row matrix is a matrix having a single row .For example

$$(a_{11} \quad a_{12} \quad a_{13})$$

is a row matrix of order 1×3 .

A column matrix is a matrix having a single column. An example of a column matrix of order 3×1 is

$$\begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix}$$

(iii) Diagonal Matrix

A square matrix in which only the leading diagonal elements exist and all the off diagonal elements are zero is called a diagonal matrix. The leading diagonal is referred to as the principle diagonal. An example of a diagonal matrix of order $m \times n$ is

$$A = \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn} \end{pmatrix}$$

A diagonal matrix can alternatively be written as

$$A = \text{diag} (a_{11}, a_{22}, a_{33}, \dots, a_{nn})$$

Another example of a diagonal matrix of order 3 is

$$A = \text{diag} (3, 6, 8)$$

Associated with the diagonal matrix are two more matrices viz. scalar matrix and unit or identity matrix

(a) Scalar matrix

A diagonal matrix whose diagonal elements are all equal is called a scalar matrix. For example

$A = \text{diag} (2, 2, 2)$ is a scalar matrix of order 3.

(b) Unit matrix or identity matrix

This is a diagonal matrix whose elements are all equal to unity. An example

$A = \text{diag} (1, 1, 1, 1)$ is a unit matrix of order 4. It can also be written as

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

(iv) Zero or null matrix

A matrix all of whose elements are zero qualifies as a null or zero matrix. It is denoted by O.

$$O = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

is a null matrix of order 3×3 .

(v) Triangular Matrix

A triangular matrix may be classified as upper triangular and lower triangular.

A square matrix A of order $n \times n$ is said to be upper triangular if $a_{ij} = 0$ for $i > j$. Thus the matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22} & a_{23} & \dots & a_{2n} \\ 0 & 0 & a_{33} & \dots & a_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & a_{nn} \end{pmatrix}$$

is a upper triangular matrix.

A square matrix A of order $n \times n$ is said to be lower triangular if $a_{ij} = 0$ for $i < j$. Thus the matrix

$$A = \begin{pmatrix} a_{11} & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & 0 & \dots & 0 \\ a_{31} & a_{32} & a_{33} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{pmatrix}$$

is a lower triangular matrix.

(vi) Symmetric matrix

A square matrix where $a_{ij} = a_{ji}$ i.e. the $(i, j)^{\text{th}}$ element = $(j, i)^{\text{th}}$ element, is called a symmetric matrix. An example of a symmetric matrix of order 3 is

$$A = \begin{pmatrix} 5 & 1 & 8 \\ 1 & 4 & 7 \\ 8 & 7 & 3 \end{pmatrix}$$

2.2.2 Matrix Operations

Addition Rules

- (i) Matrices can be added or subtracted if they are of the same order.
- (ii) The sum of two matrices result in a matrix whose elements are the sum of the corresponding elements of the corresponding matrices.

For example

$$\text{If } A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \quad \text{and}$$

$$B = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix}$$

then

$$A + B = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & a_{13} + b_{13} \\ a_{21} + b_{21} & a_{22} + b_{22} & a_{23} + b_{23} \\ a_{31} + b_{31} & a_{32} + b_{32} & a_{33} + b_{33} \end{pmatrix}$$

Rules of Matrix Addition

- (i) Commutative Rule

If A and B are two matrices of the same order, then $A + B = B + A$.

- (ii) Associative Rule

If A, B and C are three matrices of the same order, then

$$A + (B + C) = (A + B) + C$$

(iii) Distributive Rule

If A, B and C are three matrices of the same order, then

$$k(A + B) = kA + kB, \text{ k being a scalar.}$$

(iv) Existence of a Additive Identity

If A is a $m \times n$ matrix and O is a null matrix of the same order, then

$$A + O = O + A = A$$

The null matrix O is referred to as the identity of matrix addition.

(v) Existence of a Additive Inverse

For every matrix A there exists a additive inverse $-A$ such that

$$A + (-A) = O, \text{ O being the null matrix}$$

Matrix Multiplication**Scalar multiplication of a matrix**

The scalar multiplication of a matrix kA can be obtained by multiplying each and every element of A by the scalar k.

An example

If $k = 3$ and

$$A = \begin{pmatrix} 2 & 7 & 1 \\ 5 & 3 & 0 \\ 3 & 4 & 9 \end{pmatrix}$$

then

$$3A = \begin{pmatrix} 3 \times 2 & 3 \times 7 & 3 \times 1 \\ 3 \times 5 & 3 \times 3 & 3 \times 0 \\ 3 \times 3 & 3 \times 4 & 3 \times 9 \end{pmatrix} = \begin{pmatrix} 6 & 21 & 3 \\ 15 & 9 & 0 \\ 9 & 12 & 27 \end{pmatrix}$$

Product of two matrices

Two matrices A and B are said to be conformable for multiplication if the number of columns of A are equal to the number of rows of B. Thus for two matrices to be conformable for multiplication, the number of columns of the first must be equal to the number of rows of the second.

Thus if

$$A = \begin{pmatrix} 2 & 7 & 1 \\ 5 & 3 & 0 \end{pmatrix} \text{ is a } 2 \times 3 \text{ matrix}$$

and

$$B = \begin{pmatrix} 4 & 3 \\ 5 & 1 \\ 6 & -2 \end{pmatrix} \text{ is a } 3 \times 2 \text{ matrix}$$

$$\begin{aligned} \text{then } AB &= \begin{pmatrix} 2 \times 4 + 7 \times 5 + 1 \times 6 & 2 \times 3 + 7 \times 1 + 1 \times (-2) \\ 5 \times 4 + 3 \times 5 + 0 \times 6 & 5 \times 3 + 3 \times 1 + 0 \times (-2) \end{pmatrix} \text{ is a } 2 \times 2 \text{ matrix} \\ &= \begin{pmatrix} 49 & 11 \\ 35 & 18 \end{pmatrix} \end{aligned}$$

A and B are conformable for multiplication because the number of column A (3) is same as the number of rows of B (3).

Properties of Matrix Multiplication

(i) In general matrix multiplication, unlike matrix addition is not commutative.

$$AB \neq BA, \text{ in general}$$

(ii) Matrix multiplication is associative

If three matrices A, B and C are conformable for multiplication, then

$$A(BC) = (AB)C$$

The associative rule can be generalized for n matrices provided the products are defined

$$A_1 A_2 (A_3 A_4 \dots \dots A_n) = A_1 (A_2 A_3 A_4 \dots \dots) A_n = A_1 (A_2 A_3 A_4 \dots \dots A_n) = \dots \dots$$

(iii) Matrix multiplication is distributive with respect to addition

This means for three matrices A, B and C

$$A(B+C) = AB + AC$$

$$(B+C)D = BD + CD$$

provided of course that the matrices are conformable for the operations defined.

(iv) If A is a $m \times n$ matrix and O is a null matrix of order $n \times m$ then $AO = OA$.

If A is a square matrix of order n and I is a unit matrix of the same order then

$$AI = IA = A.$$

2.2.3 Applications of Matrices in Business Problems

We now examine simple applications of matrix theory in business situations. These examples highlight the ease with which data can be arranged in the form of matrices and thus facilitate easy analysis of decision making situations.

Example 2.16: A manufacturing company has production facilities at two locations A and B. Each facility produces two different types of Air Conditioners (AC) in two different categories I and II. The production figures (in '000's) at location A are as follows.

	Type I	Type II
Category I	55	70
Category II	60	80

At Plant B the production figures are:

	Type I	Type II
Category I	50	80
Category II	85	70

Find the total number of AC's produced in both plants.

Solution:

Matrix addition can now be used to find the total production of AC's across the two plants and across the different categories. This will consist of the simple step of considering the first plant data as matrix A and the second plant data as matrix B and then adding the two matrices A and B as follows.

	Type I	Type II
Category I	105	150
Category II	145	150

The management is now contemplating opening a third plant with thrice the capacity of the first plant. Scalar multiplication of matrices can be used to represent the production at the third plant.

	Type I	Type II
Category I	$3 \times 55 = 165$	$3 \times 70 = 210$
Category II	$3 \times 60 = 180$	$3 \times 80 = 240$

Example 2.17: An insurance company has set up offices at 9 capital cities of major states and 4 metros Delhi, Kolkata, Mumbai and Chennai. Each office has primarily three posts viz. manager executives and assistant. The staff composition for the different office locations are

	Manager	Executives	Assistants
Capital cities	1	2	1
Metros	2	4	5

The basic daily salaries for the different positions are as follows

Manager: Rs. 1500

Executive: Rs. 800

Assistants: Rs. 300

Use matrices to determine the total number of posts of each kind in all the offices, the total salary born by the company each day for each office and the total salary of all the offices taken together.

Solution:

First of all putting all this information in the form of matrices we have the number of offices as

A: Offices (a 1×2 row matrix)

$$A = \begin{pmatrix} \text{Capitals} & \text{Metros} \\ 9 & 4 \end{pmatrix}$$

B: Staff composition (a 2×3 matrix)

$$\begin{matrix} \text{Capital cities} \\ \text{Metros} \end{matrix} \begin{pmatrix} \text{Manager} & \text{Executives} & \text{Assistants} \\ 1 & 2 & 1 \\ 2 & 4 & 5 \end{pmatrix}$$

C: Basic salaries (a column matrix)

$$\begin{matrix} \text{Manager} \\ \text{Executive} \\ \text{Assistants} \end{matrix} \begin{pmatrix} \text{Rs. 1500} \\ \text{Rs. 800} \\ \text{Rs. 300} \end{pmatrix}$$

The matrix multiplication AB will result in the total number of posts in all the offices.

$$\text{Total Posts} \begin{pmatrix} \text{Manager} & \text{Executives} & \text{Assistants} \\ 9+8=17 & 18+16=34 & 9+20=29 \end{pmatrix}$$

The salary requirement of each office located at capital cities and metros can be obtained by the product $D = BC$

$$\begin{matrix} \text{Capital cities} \\ \text{Metros} \end{matrix} \begin{pmatrix} \text{Salaries} \\ 1500 + 1600 + 300 = 3400 \\ 3000 + 3200 + 1500 = 7700 \end{pmatrix}$$

Finally the total salaries of all the offices taken together can be obtained by AD

$$AD = (9 \ 4) \begin{pmatrix} 3400 \\ 7700 \end{pmatrix} \\ = \text{Rs. 61,400}$$

Example 2.18: A manufacturer produces 3 models M_1, M_2, M_3 of his product. Each model contains 4 types of sub-assemblies, A_1, A_2, A_3, A_4 . The number of sub-assemblies of each type contained in each type contained in each model is given by the matrix:

$$\begin{matrix} \text{Sub} \\ M_1 \\ M_2 \\ M_3 \end{matrix} \begin{pmatrix} \text{assemblies} \\ A_1 & A_2 & A_3 & A_4 \\ 6 & 3 & 8 & 12 \\ 8 & 7 & 0 & 3 \\ 15 & 14 & 5 & 6 \end{pmatrix}$$

Each sub - assembly contains basic parts of type B_1 and B_2 and their number is given by the following matrix:

Basic Parts:

$$\begin{matrix} & B_1 & B_2 \\ \begin{pmatrix} 20 & 14 \\ 17 & 21 \\ 9 & 25 \\ 6 & 4 \end{pmatrix} \end{matrix}$$

Find the number of basic parts of each type necessary for each type of model of the product. If B_1 and B_2 cost Rs.25 and Rs.20 per unit respectively, find the cost of the basic parts that constitute each model.

Solution:

The number of basic parts of type B_1 and B_2 is given by multiplying the two matrices as follows:

$$AB = \begin{pmatrix} 6 & 3 & 8 & 12 \\ 8 & 7 & 0 & 3 \\ 15 & 14 & 5 & 6 \end{pmatrix} \begin{pmatrix} 20 & 14 \\ 17 & 21 \\ 9 & 25 \\ 6 & 4 \end{pmatrix}$$

$$\begin{matrix} & B_1 & B_2 \\ \begin{matrix} M_1 \\ M_2 \\ M_3 \end{matrix} \begin{pmatrix} 6 \times 20 + 3 \times 17 + 8 \times 9 + 12 \times 6 & 6 \times 14 + 3 \times 21 + 8 \times 9 + 12 \times 6 \\ 8 \times 20 + 7 \times 17 + 0 \times 9 + 3 \times 6 & 8 \times 14 + 7 \times 21 + 0 \times 25 + 4 \times 3 \\ 15 \times 20 + 14 \times 17 + 5 \times 9 + 6 \times 6 & 15 \times 14 + 14 \times 21 + 5 \times 25 + 6 \times 4 \end{pmatrix} \end{matrix}$$

$$\begin{matrix} & B_1 & B_2 \\ \begin{matrix} M_1 \\ M_2 \\ M_3 \end{matrix} \begin{pmatrix} 120 + 51 + 72 + 72 & 84 + 63 + 200 + 48 \\ 160 + 119 + 0 + 18 & 112 + 147 + 12 \\ 300 + 238 + 45 + 36 & 210 + 294 + 125 + 24 \end{pmatrix} \end{matrix}$$

$$\begin{matrix} & B_1 & B_2 \\ \begin{matrix} M_1 \\ M_2 \\ M_3 \end{matrix} \begin{pmatrix} 315 & 395 \\ 297 & 271 \\ 619 & 653 \end{pmatrix} \end{matrix}$$

Since B_1 cost Rs.25 and B_2 costs Rs.20 per unit the cost of the basic parts for each model is calculated as follows:

$$\text{Model } M_1 : 25 \times 315 + 20 \times 395 = 7875 + 7900 = \text{Rs. } 15775$$

$$\text{Model } M_2 : 25 \times 297 + 20 \times 271 = 7425 + 5450 = \text{Rs. } 12845$$

$$\text{Model } M_3 : 25 \times 619 + 20 \times 653 = 15475 + 13060 = \text{Rs. } 28535$$

Example 2.19: The annual sales volumes of 3 products A, B & C whose sales prices per unit are Rs.4, Rs.2 & Rs.1 respectively in two different cities are given below in a matrix form.

Cities	Products		
	A	B	C
X	5,000	8,000	12,000
Y	11,000	5,000	16,000

Find the total revenue generated in each market using matrices.

Solution:

The price matrix for the three products A, B and C is:

$$P = \begin{pmatrix} A & 4 \\ B & 2 \\ C & 1 \end{pmatrix}$$

The volume matrix is:

$$V = \begin{pmatrix} 5,000 & 8,000 & 12,000 \\ 11,000 & 5,000 & 16,000 \end{pmatrix}$$

The revenue matrix is:

$$\begin{aligned} VP &= \begin{pmatrix} 5,000 & 8,000 & 12,000 \\ 11,000 & 5,000 & 16,000 \end{pmatrix} \begin{pmatrix} 4 \\ 2 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 5000 \times 4 + 8000 \times 2 + 12000 \\ 11,000 \times 4 + 5000 \times 2 + 16000 \end{pmatrix} = \begin{pmatrix} 48000 \\ 70000 \end{pmatrix} \end{aligned}$$

The total revenue in the first city is Rs. 48,000.

The total revenue in second city is Rs. 70,000.

Example 2.20: A manufacturer produces three products: P, Q and R, which he sells in two markets. Annual sale volumes are indicated as follows:

Markets	Products		
	P	Q	R
I	10,000	2,000	18,000
II	6,000	20,000	8,000

1. If the unit sale prices of P, Q and R are Rs. 3.00, Rs. 2.00, and Rs. 1.00 respectively, find the total revenue in each market with the help of matrix algebra.
2. If the unit costs of the above 3 commodities are Rs. 2.00, Rs. 1.50 and Re.0.80 respectively, find the gross profit.

Solution:

- Volume Matrix: $V = \begin{pmatrix} 10,000 & 2,000 & 18,000 \\ 6,000 & 20,000 & 8,000 \end{pmatrix}$

- Sales matrix: $S = \begin{matrix} P & \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix} \\ Q \\ R \end{matrix}$

The total revenue matrix is given by the matrix multiplication.

$$\begin{aligned} VS &= \begin{pmatrix} 10,000 & 2,000 & 18,000 \\ 6,000 & 20,000 & 8,000 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 30,000 + 4,000 + 18,000 \\ 18,000 + 40,000 + 8,000 \end{pmatrix} \\ &= \begin{pmatrix} 52,000 \\ 66,000 \end{pmatrix} \end{aligned}$$

Thus, the revenue generated by the sales in the first market is Rs.52,000 and the revenue generated by sales in the second market is Rs.66,000.

- The Cost Matrix

$$C = \begin{pmatrix} 2.00 \\ 1.50 \\ 0.80 \end{pmatrix}$$

The total cost

$$\begin{aligned} VC &= \begin{pmatrix} 10,000 & 2,000 & 18,000 \\ 6,000 & 20,000 & 8,000 \end{pmatrix} \begin{pmatrix} 2.00 \\ 1.50 \\ 0.80 \end{pmatrix} \\ &= \begin{pmatrix} 20,000 + 3,000 + 14,400 \\ 12,000 + 30,000 + 6,400 \end{pmatrix} \\ &= \begin{pmatrix} 37,400 \\ 48,400 \end{pmatrix} \end{aligned}$$

$$\begin{aligned} \text{The profit is } VS - VC &= \begin{pmatrix} 52,000 \\ 66,000 \end{pmatrix} - \begin{pmatrix} 37,400 \\ 48,400 \end{pmatrix} \\ &= \begin{pmatrix} 14,600 \\ 17,600 \end{pmatrix} \end{aligned}$$

The profit from the first market = Rs. 14,600

The profit from the second market = Rs. 17,600

Example 2.21: One thousand airline stewardess, classified in the following groups, are employed by an airline:

Age:	20-24	25-29	30-34	35-39
Number:	600	300	150	30

The probability that a stewardess of the corresponding group will leave by the end of the year is given by the column matrix.

$$A = \begin{pmatrix} 0.45 \\ 0.30 \\ 0.15 \\ 0.10 \end{pmatrix}$$

Using matrix multiplication, find the number of stewardesses who are expected to leave in one year.

Solution:

Number of employees or Employee matrix

$$E = (600 \ 300 \ 150 \ 30)$$

The probability matrix

$$A = \begin{pmatrix} 0.45 \\ 0.30 \\ 0.15 \\ 0.10 \end{pmatrix}$$

Thus the number of stewardesses who are expected to leave in a year.

$$EA = (600 \ 300 \ 150 \ 30) \begin{pmatrix} 0.45 \\ 0.30 \\ 0.15 \\ 0.10 \end{pmatrix}$$

$$= (270 + 90 + 22.5 + 3)$$

$$= 385.5$$

i.e. is approximately 385 stewardesses

2.3 DIFFERENTIAL CALCULUS: APPLICATIONS

Differential calculus has a wide variety of business applications. We are going to explore in brief its application with respect to it being considered as a rate of change. Rate of change of a

function is applicable in everyday life. For example rate of change of prices, costs, speed etc. In our discussion on linear problems we discussed two variables, the dependent or influenced variable and the independent or influencing variable. In this section, we discuss the rate of variation in the dependent variable with respect to infinitesimal changes in the independent variable. For example for a particular demand function, it would be possible to find the degree of change in demand with reference to a small change in price or income or perhaps both as the case may be. Also typical applications of the maxima and minima principles of differential calculus in terms of profit maximization and cost minimization will be discussed at the end of this chapter.

2.3.1 Overview of Differential Calculus

Let $y = f(x)$ be a function. A simple definition of derivative is the instantaneous rate of change of a variable y (dependent) with respect to another variable x (independent). It is denoted by

$$\frac{dy}{dx}$$

General rules of differentiation

Rule 1. Let $y = x^n$

$$\frac{dy}{dx} = \frac{dx^n}{dx} = n x^{n-1}$$

Example 2.22: If $y = x^2$ then $\frac{dy}{dx} = 2x$

Example 2.23: If $y = x^6$ then $\frac{dy}{dx} = 6x^5$

Rule 2. The derivative of the product of a constant and a function is the product of the constant and the derivative of the function.

$$\frac{d(ku)}{dx} = k \frac{du}{dx}$$

Example 2.24: If $y = 10x^3$

$$\text{Then } \frac{dy}{dx} = 10 \left(\frac{dx^3}{dx} \right) = 10 \times 3x^2 = 30 x^2$$

Rule 3. The derivative of a sum of finite number of functions is the sum of their derivatives.

$$\begin{aligned} & \frac{d}{dx} [f(x) + \phi(x) + \dots\dots\dots] \\ &= \frac{df(x)}{dx} + \frac{d\phi(x)}{dx} + \dots\dots \end{aligned}$$

Example 2.25: If $y = 3x + 4 x^2$ then

$$\frac{dy}{dx} = 3 + 4 \frac{dx^2}{dx} = 3 + 8x$$

Rule 4: The derivative of the product of two functions is equal to the first function into the derivative of the second plus the second function into the derivative of the first.

$$\frac{d}{dx}[f(x)g(x)] = f(x)\frac{d}{dx}g(x) + g(x)\frac{d}{dx}f(x)$$

Rule 5: The derivative of any constant is zero.

$$\frac{d}{dx}c = 0, \text{ if } c \text{ is a constant}$$

Example 2.26: If $y = (3x + 2)(2x^2 + 4x + 5)$

Then

$$\begin{aligned} \frac{dy}{dx} &= (3x + 2) \left(2 \frac{dx^2}{dx} + 4 \frac{dx}{dx} + \frac{d5}{dx} \right) + (2x^2 + 4x + 5) 3 \frac{dx}{dx} \\ &= (3x + 2)(4x + 4 + 0) + 3(2x^2 + 4x + 5) \\ &= 12x^2 + 12x + 8x + 8 + 6x^2 + 12x + 15 \\ &= 18x^2 + 32x + 23 \end{aligned}$$

Rule 5: (Quotient Rule) If f and g are differentiable functions and $g(x) \neq 0$ then

$$\frac{d}{dx} \left[\frac{f(x)}{g(x)} \right] = \frac{g(x)f'(x) - f(x)g'(x)}{[g(x)]^2}$$

i.e., the derivative of the quotient of two functions is the denominator times the derivative of the numerator, minus the numerator times the derivative of the denominator, all divided by the square of the denominator.

Example 2.27: Let $y = \frac{x^3 + 1}{x^2 - 1}$. Find $\frac{dy}{dx}$

Solution:

$$y = \frac{x^3 + 1}{x^2 - 1}$$

$$\frac{dy}{dx} = \frac{(x^2 - 1)(3x^2) - (x^3 + 1)(2x)}{(x^2 - 1)^2}, \text{ By applying the Quotient Rule}$$

$$= \frac{3x^4 - 3x^2 - 2x^4 - 2x}{(x^2 - 1)^2}$$

$$= \frac{x^4 - 3x^2 - 2x}{(x^2 - 1)^2}$$

2.3.2 Applications of Derivative

2.3.2.1 Price Elasticity of Demand

The price elasticity of demand is the ratio of the proportionate change in quantity demanded by the proportionate change in price.

Mathematically if $p = f(x)$ is the demand function (p being the price & x the quantity demanded).

Then

$$p_d = - \frac{p}{x} \frac{dx}{dp}$$

where P_d - price elasticity of demand

Interpretation

when $p_d < 1$, demand is elastic

$p_d > 1$, demand is inelastic

$p_d = 1$, demand is unitary

2.3.2.2 Price elasticity of Supply

Price elasticity of supply is the ratio of the proportional change in quantity supplied by the proportionate change in price.

Mathematically, if $p = g(x)$ is the supply function (p being the price and x the quantity supplied)

Then
$$p_s = \frac{p}{x} \frac{dx}{dp}$$

Example 2.28: Given the demand function $p = \sqrt{50 - x^2}$ and the supply function $x = 3p - 20$ where p is price and q is quantity.

Find the elasticity of demand and supply at the equilibrium price.

Solution:

The demand function is

$$p = \sqrt{50 - x^2}$$

$$p^2 = 50 - x^2$$

$$x = \sqrt{50 - p^2}$$

...(1)

Also supply function x is given as $x = 3p - 20$

... (2)

The equilibrium price can be obtained by equating 1 & 2

$$\sqrt{50 - p^2} = 3p - 20$$

$$\Rightarrow 50 - p^2 = 9p^2 - 120p + 400$$

$$\Rightarrow 10p^2 - 120p + 350 = 0$$

$$\Rightarrow p^2 - 12p + 35 = 0$$

$$\Rightarrow p^2 - 7p - 5p + 35 = 0$$

$$\Rightarrow (p - 7)(p - 5) = 0$$

$$\Rightarrow p = 5, 7$$

Thus there are equilibrium prices at Rs.5 & Rs.7. We now find the elasticities at the price Rs.5. The calculations for Rs. 7 will be similar.

Price elasticity of demand

$$\begin{aligned}
 P_d &= -\frac{p}{x} \frac{dx}{dp} \\
 &= -\frac{p}{x} \frac{d}{dp} (50 - p^2)^{1/2} \\
 &= \frac{-p}{x} \frac{1}{2} (50 - p^2)^{-1/2} (-2p) \\
 &= \frac{p}{x} \frac{p}{\sqrt{50 - p^2}} \\
 &= \frac{p^2}{\sqrt{50 - p^2} \sqrt{50 - p^2}}
 \end{aligned}$$

Price elasticity of supply

$$\begin{aligned}
 P_s &= \frac{p}{x} \frac{dx}{dp} \\
 &= \frac{p}{x} \frac{d}{dp} (3p - 20) \\
 &= \frac{p}{x} 3 = \frac{3p}{x} = \frac{3p}{3p - 20}
 \end{aligned}$$

when $p = \text{Rs.}5$

$$P_d = \frac{5^2}{50 - 5^2} = \frac{25}{50 - 25} = \frac{25}{25} = 1$$

$$P_s = \frac{3 \times 5}{3 \times 5 - 20} = \frac{15}{-5} = -3$$

2.3.2.3 Average Cost and Marginal Cost

If C is the total cost of production of x units of a commodity and

$$C = f(x)$$

Then the average cost AC is

$$AC = \frac{C}{x} = \frac{f(x)}{x}$$

The marginal cost is the rate of change in C corresponding to a unit change in x , Thus

$$MC = \frac{dC}{dx}$$

It may thus be defined as the approximate cost of one additional unit of output.

Example 2.29: The total cost $C(x)$ of a firm is given by: $C(x) = 20 + 2x + 0.5x^2$ in terms of the output x . Determine (i) average cost (AC), (ii) slope of AC, (iii) marginal cost (MC), (iv) slope of MC.

Solution:

$$1. \text{ Average cost (AC)} = \frac{C(x)}{x} = \frac{20 + 2x + 0.5x^2}{x} = \frac{20}{x} + 2 + 0.5x$$

$$2. \text{ Slope of Average Cost} = \frac{d}{dx}(\text{AC}) = -\frac{20}{x^2} + 0.5$$

$$3. \text{ The Marginal Cost (MC): } \frac{dc}{dx} = \frac{d}{dx}(20 + 2x + 0.5x^2)$$

$$4. \text{ Slope of MC} = \frac{d}{dx}(\text{MC}) = \frac{d}{dx}(2 + x) = 1$$

2.3.2.4 Maxima & Minima

A very important application of the theory of calculus is in determining the maxima and minima of functions. For example a manufacturer may want to know the number of units of the product to be produced to minimize the average cost. Alternately the manufacturer may also want to evaluate the number of units of the product to be produced so as to maximize his/her revenue generation. These questions can be answered with the application of the theory of maxima and minima in differential calculus.

Let us first look mathematically at this situation and then interpret it with the help of some examples.

For both maxima and minima, two conditions need to be fulfilled. If $y = f(x)$ is a function, then

For maximum, the first order condition is $\frac{dy}{dx} = 0$ and

Second order condition is $\frac{d^2 y}{dx^2} < 0$

For a function to achieve a minima, the first order condition is $\frac{dy}{dx} = 0$ and

Second order condition is $\frac{d^2 y}{dx^2} > 0$

This is summarized in the following table

Table 2.1
Maxima and Minima Condition

	First order condition	Second order condition
Maxima	$\frac{dy}{dx} = 0$	$\frac{d^2 y}{dx^2} < 0$
Minima	$\frac{dy}{dx} = 0$	$\frac{d^2 y}{dx^2} > 0$

Example 2.30 A company produces x units of output at a total cost of Rs. $(x^3 - 78x^2 + 2500x)$. Find:

1. Output at which marginal cost is minimum.
2. Output at which average cost is minimum.
3. Output at which marginal cost is equal to the average cost.

Solution:

Total cost function: $x^3 - 78x^2 + 2500x$

Marginal cost function: $3x^2 - 156x + 2500$ (MC)

Average cost function: $x^2 - 78x + 2500$ (AC)

1. For minima: $\frac{d}{dx}(\text{MC}) = 0$

$$\Rightarrow 6x - 156 = 0$$

$$\Rightarrow x = 26$$

Also $\frac{d^2}{dx^2}(\text{MC}) = 6 > 0$

Thus the output at which marginal cost is minimum is 26 units.

2. For average cost minima

$$\frac{d}{dx}(\text{AC}) = 0$$

$$\Rightarrow 2x - 78 = 0$$

$$x = 39$$

$$\frac{d^2}{dx^2}(\text{AC}) = 2 > 0$$

Thus output at which average cost is minimum is 39 units.

3. Setting the average cost equal to the marginal cost

$$\Rightarrow 3x^2 - 156x + 2500 = x^2 - 78x^2 + 2500$$

$$\Rightarrow 2x^2 - 78x = 0$$

$$\Rightarrow x(2x - 78) = 0$$

$$\Rightarrow x = 0, x = 39$$

Thus the output at which marginal cost is equal to the average cost is 39 units.

Example 2.31: Determine the maximum and minimum values of the function.

$$y = x^3 - 3x^2 + 3x$$

Solution:

First order condition

$$\frac{dy}{dx} = 3x^2 - 6x + 3 = 0$$

$$\Rightarrow x^2 - 2x + 1 = 0$$

$$\Rightarrow x^2 - x - x + 1 = 0$$

$$\Rightarrow x(x-1) - 1(x-1) = 0$$

$$\Rightarrow (x-1)^2 = 0$$

$$\Rightarrow x = 1$$

Second order condition

For maximum:

$$\frac{d^2 y}{dx^2} < 0$$

$$\frac{d^2 y}{dx^2} = 6x - 6$$

$$\text{At } x = 1$$

$$\frac{d^2 y}{dx^2} = 6 - 6$$

$$= 0$$

A value of 0 indicates that nothing can be said about the maximum or minimum value of the function.

Example 2.32: Determine the maximum & minimum values of the function

$$y = \frac{2}{3}x^3 + \frac{1}{2}x^2 - 6x + 8$$

Solution:

First order condition

$$\frac{dy}{dx} = 2x^2 + x - 6 = 0$$

$$\Rightarrow (x + 2)(2x - 3) = 0$$

$$\Rightarrow \frac{dy}{dx} = 0 \text{ at } x = -2, \frac{3}{2}$$

Second order condition:

$$\frac{d^2 y}{dx^2} = 4x + 1$$

For maxima, we have to check whether

$$\frac{d^2 y}{dx^2} < 0 \text{ at } x = -2 \text{ or } x = \frac{3}{2}$$

At $x = -2$

$$\frac{d^2 y}{dx^2} = 4(-2) + 1 = -8 + 1 = -7 < 0$$

Thus, the function attains a maxima at $x = -2$.

For minimima, we have to check whether

$$\frac{d^2 y}{dx^2} > 0 \text{ at } x = \frac{3}{2}$$

$$\frac{d^2 y}{dx^2} = 4\left(\frac{3}{2}\right) + 1 = 7 > 0$$

Thus, the given function attains a minima at $x = \frac{3}{2}$

Example 2.33: A firm manufacturing pens produces an output of x units at a total variable cost given by

$$C = x^3 - 4x^2 + 7x$$

Find the average variable cost and also the output at which average variable cost is minimum and the output at which the average variable cost is maximum.

Solution:

Average variable cost

$$AVC = \frac{c}{x} = \frac{x^3 - 4x^2 + 7x}{x}$$

$$\text{Or } y = x^2 - 4x + 7$$

To find the maximum and minimum outputs, we now have to verify the first and second order conditions

First order condition:

$$\frac{dy}{dx} = 2x - 4 = 0$$

$$x = 2$$

Second order condition:

$$\frac{d^2 y}{dx^2} = 2 > 0$$

Thus the firm attains a maximum average variable cost at 2 units.

Example 2.34: A shoe manufacturer has computed his cost function to be

$$C(x) = x^2 - 1200x + 3,60,040$$

and his revenue function as $R(x) = 9750x - 75x^2$ where x is the number of pairs of shoes. Find how many pairs of shoes have to be produced to maximize revenue and how many pairs will minimize his cost.

Solution:

Cost function: $C(x) = x^2 - 1200x + 3,60,040$

Revenue function: $R(x) = 9750x - 75x^2$

We first find out at what output, cost will be minimum.

First order condition:

$$\frac{dC(x)}{dx} = 2x - 1200 = 0$$

$$x - 600 = 0$$

$$x = 600$$

Second order condition:

$$\frac{d^2 y}{dx^2} = 2 > 0$$

Thus the cost for the manufacturer is minimum at 600 units.

We now calculate the output at which his revenue will be maximum.

$$\frac{dR(x)}{dx} = -150x + 9750 = 0 \Rightarrow x = 65$$

$$\frac{d^2 R(x)}{dx^2} = -150 < 0$$

Thus, the manufacturer would have to produce 65 pairs of shoes to maximize his revenue.

Example 2.35: A company charges Rs.1000 for hiring computers on orders of Rs.50 or less. The charge is reduced by Rs.10 per set for each set ordered in excess of 50. Find the largest size order the company should allow to get maximum revenue.

Solution:

Let x be the number of computers ordered above 50. Total number of computers ordered = $50 + x$. For computers above 50, the charges become Rs. $(1000 - 10x)$.

The total revenue function

$$\begin{aligned} R &= (50 + x)(1000 - 10x) \\ &= 50,000 - 500x + 1000x - 10x^2 \\ R &= -10x^2 + 500x + 50000 = 0 \end{aligned}$$

To maximize the revenue

$$\frac{dR}{dx} = 0 \text{ \& } \frac{d^2R}{dx^2} < 0$$

$$\frac{dR}{dx} = 0$$

$$\Rightarrow 20x = 500$$

$$\Rightarrow x = 25$$

$$\frac{d^2R}{dx^2} = -20 < 0$$

Thus, the largest size order is $50 + 25 = 75$ computers to get maximum revenue.

Example 2.36:

- (a) A given product can be manufactured at a total cost $C(x) = \text{Rs.} \left(\frac{x^2}{100} + 100x + 40 \right)$, where x is the number of units produced. The price at which each unit can be sold is given by: $p = \text{Rs.} \left(200 - \frac{x}{400} \right)$. Determine the production level x at which the profit is maximum. What is the price per unit and the total profit at this level of production?
- (b) A firm finds that it can sell all that it produces (within limits). The demand function is $p = 260 - 3x$, where p is the price per unit at which it can sell x units. The cost function is $C = 500 + 20x$, where x is the number of units produced. Find x so that the profit is maximum.

Solution:

(a) Total cost: $C(x) = \frac{x^2}{100} + 100x + 40$

$$\begin{aligned} \text{Revenue function: } R(x) &= px \\ &= \left(200 - \frac{x}{400} \right) x \end{aligned}$$

$$\begin{aligned} \text{Profit: } p(x) &= 200x - \frac{x^2}{400} - \left(\frac{x^2}{100} + 100x + 40 \right) \\ &= 200x - \frac{x^2}{400} - \frac{x^2}{100} - 100x - 40 \end{aligned}$$

$$\begin{aligned}
 &= 200x - \left(\frac{x^2 + 4x^2}{400} \right) - 100x - 40 \\
 &= 100x - \frac{5x^2}{400} - 40 \\
 &= \frac{-x^2}{80} + 100x - 40
 \end{aligned}$$

The production level at which the profit is maximum is obtained from the condition:

$$\begin{aligned}
 \frac{dp(x)}{dx} &= 0 \\
 \Rightarrow \frac{-2x}{80} &= -100 \\
 \Rightarrow x &= 400 \text{ units}
 \end{aligned}$$

The price per unit = $200 - \frac{400}{400} = \text{Rs.}199$

The total profit at this level of production

$$\begin{aligned}
 &= \frac{-400^2}{80} + 400 \times 100 - 40 \\
 &= -2000 + 40000 - 40 \\
 &= \text{Rs.}37960
 \end{aligned}$$

(b) Demand fn: $p = 260 - 3x$

$$\Rightarrow x = \frac{260 - p}{3}$$

$$\begin{aligned}
 \text{Profit fn: } p(x) &= xp - (500 + 20x) \\
 &= x(260 - 3x) - (500 + 20x) \\
 &= 260x - 3x^2 - 500 - 20x \\
 &= -3x^2 + 240x - 500
 \end{aligned}$$

To maximize profit

$$\begin{aligned}
 \frac{dp(x)}{dx} &= 0 \ \& \ \frac{d^2p(x)}{dx^2} < 0 \\
 \frac{dp(x)}{dx} = 0 &\Rightarrow -6x + 240 = 0 \\
 \Rightarrow x &= 40
 \end{aligned}$$

$$\frac{d^2p(x)}{dx^2} = -6 < 0$$

Thus profit is maximum when number of units produced is 40.

Example 2.37: A manufacturer can sell x items per day at a price p rupees each, where, the cost of production for x items is $500 + 13x + 0.2x^2$. $p = 125 - \frac{5}{3}x$.

1. Find the volume, which will give maximum profit.
2. What is the maximum profit?

Solution:

$$1. \text{ Profit: } \left(125 - \frac{5}{3}x\right)x - (500 + 13x + 0.2x^2)$$

$$\Rightarrow p = 125x - \frac{5}{3}x^2 - 500 - 13x - 0.2x^2$$

$$\Rightarrow p = -1.87x^2 + 112x - 500$$

The volume, which will maximize profit, is given by the twin conditions

$$\frac{dp}{dx} = 0 \text{ \& } \frac{d^2p}{dx^2} < 0$$

$$\frac{dp}{dx} = 0 \Rightarrow -3.74x + 112 = 0$$

$$\Rightarrow x \cong 30 \text{ units of the item would maximize profit.}$$

2. The maximum profit can be obtained by putting $x = 30$ in the profit function.

$$\begin{aligned} P &= -1.87(30)^2 + 112(30) - 500 \\ &= -1683 + 3360 - 500 \\ &= \text{Rs. } 1177 \end{aligned}$$

2.4 EXERCISES

- 2.1 A calculator manufacturer finds that the production costs directly attributable to each calculator is Rs.20 and the fixed costs are Rs.10, 000. If each calculator can be sold for Rs.30,
 - (i) Determine the cost function,
 - (ii) Determine the revenue function, and
 - (iii) Determine the break-even point
- 2.2 A belt manufacturer determines that the production costs associated with each belt are Rs.25 and the fixed costs are Rs.10, 000. Assuming that each belt produced can be sold for Rs.50, determine the break – even point.
- 2.3 A company decides to set up a small production plant for manufacturing wrist watches. The cost for initial set up is Rs.10 lakhs. The additional cost for producing each watch is Rs.500. Each watch is sold at Rs.1000. During the first month 1,500 watches are produced and sold:
 - (i) Determine the total cost function $C(x)$ for the production of x watches.

- (ii) Determine the revenue function $R(x)$.
- (iii) Determine the profit function $P(x)$.
- (iv) How much profit or loss the company incurs during the first month when all the 1,500 watches are sold?
- (v) Determine the break – even point.
- 2.4 The cost of producing x items is given by $C(x) = 2.80x + 800$ and each item sells for Rs.5.
- (i) Find the break – even point.
- (ii) If it is known that at least 500 units will be sold, what should be the price charged for each item to guarantee no loss?
- 2.5 A computer manufacturer determines that its total cost for x number of sets is given by $C(x) = 500x^2 + 2500x + 5000$. Each set sells for Rs.10, 000. Determine the break – even point.
- 2.6 An automobile spare part manufacturing company introduces production bonus to the employees that increases the cost of the spare part. The daily cost of production C for x number of spare parts is given by:
- $C(x) = \text{Rs.}2.50x + \text{Rs.}550$
- (i) If each spare part is sold for Rs.3, determine the minimum number that must be produced and sold daily to ensure no loss.
- (ii) If the selling price is increased by 30 paise per piece, what would be the break – even point?
- (iii) If it is known that at least 500 parts can be sold daily, what price the company should charge per piece of spare part to guarantee no loss?
- 2.7 A shopkeeper earns Rs.380 in the first week, Rs.660 in the second week and Rs.860 in the third week. On plotting the points (1,380), (2,660) and (3,860), the shopkeeper feels that a quadratic function may fit the data.
- (i) Find the quadratic function that fits the data.
- (ii) Using your model make a prediction of the earning for the fourth week.
(From past C.A. examination, May 3, 1995)
- 2.8 The cost function $C(x)$ for 'x' breads is given by:
- $C(x) = \text{Rs.}3.5x + \text{Rs.}12,000$
- Each bread is put to a special levy of 20 paise for Andhra Pradesh cyclone victims. Then
- (i) If each bread is sold for Rs.6, determine the minimum number of breads that should be produced and sold to ensure no loss.
- (ii) If the selling price is increased by 70 paise per bread, what would be the break – even point?
- (iii) If 6,000 breads are sold only, what price per bread should be charged to guarantee no loss?
(From past C.A. examination, May 5, 1997)
- 2.9 A company manufacturers wrist watches for which $P = 1500 - 3x$, represents the demand function; where p is the price per unit of x units. Cost price involves initially a fixed cost of Rs.38, 400 and a variable cost of rs.420 per watch. Find at what level of production company expects to recover its cost.
(From past C.A. examination, Nov 3, 1997)

2.10 A cottage toy industry has 29 workers. The cost of producing a unit of toy is Rs.2.07. offer fixed price cost including production bonus is Rs.30 per worker.

- (i) If each toy is sold for Rs.6, determine the number of toys that must be produced and sold daily to ensure no loss.
- (ii) If to promote sale, price is reduced by 50 paise per toy, what would be break – even point and if at this rate 500 toys are sold daily, what would be the profit?

(From past C.A. examination, May 4, 1999)

$$2.11 \text{ Given } A = \begin{pmatrix} 1 & 4 & 7 \\ 6 & 5 & 8 \\ 2 & 4 & 6 \end{pmatrix}, \quad B = \begin{pmatrix} 8 & 4 & -3 \\ 3 & -2 & 0 \\ 2 & 6 & 2 \end{pmatrix}, \quad C = \begin{pmatrix} 8 & 2 & 0 \\ -8 & 4 & -10 \\ 0 & 2 & -6 \end{pmatrix}$$

1 Compute the following:

- (i) $A + B$; (ii) $A - B$; (iii) $A + (B + C)$; (iv) $(A + B) + C$;
 (v) $(A-B) + C$; (vi) $A-B-C$; (vii) $2(A+B)$; (viii) $2A + 2B$
 (ix) $3A + 2B - 3C$; (x) $3B + 2A$; (xi) $2B + 3A$

$$2.12 \text{ } A = \begin{pmatrix} 1 & 3 \\ 3 & 6 \\ 5 & 8 \end{pmatrix} \quad B = \begin{pmatrix} 3 & 5 & 9 \\ 6 & -2 & 1 \end{pmatrix}$$

- (i) Write down the order of the matrices A and B.
 (ii) Write down the order of the product AB.
 (iii) Calculate AB
 (iv) Is it possible to calculate BA?
 (v) Is $AB = BA$
 (vi) Are the following possible? $A + B$, $A - B$, $2B - A^2$

$$2.13 \text{ } A = \begin{pmatrix} 1 & 2 \\ 4 & 5 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 0 \\ -3 & 4 \end{pmatrix} \quad C = \begin{pmatrix} 1 & - \\ 1 & 0 \end{pmatrix} \text{ show that:}$$

- (i) $A(B + C) = AB + AC$
 (ii) $(AB)C = A(BC)$

2.14 A company is marketing 4 different types of mobile phones. Although the four models have the same rating, the principal difference between them lies in the combination of accessories. For example, one type may not have built in camera and another may be without GPRS. Five parts are required in various quantities depending upon the model and the following tabulations shows the requirements:

Phone Model	Parts Required				
	A	B	C	D	E
I	1	2	0	5	2
II	0	3	0	1	5
III	1	1	4	2	2
IV	1	2	4	5	5

- (i) What will be the requirements of the parts A,B, C,D,E if the company has to supply 3 model I phones, 5 model II phones, 2 model III phones, and 10 model IV phones?
- (ii) If the cost of the parts A, B, C, D, E be Rs.30, Rs.12, Rs.5, Rs.4 and Rs.7 respectively, find the amount spent on purchasing all the parts of these phones.

2.15 Two shops have in stock large, medium and small sizes of a brand shampoo. The number of each size stocked is given by the matrix A, where

Large Medium Small

$$A = \begin{pmatrix} 150 & 240 & 120 \\ 90 & 300 & 210 \end{pmatrix} \begin{matrix} \text{Shop No. 1} \\ \text{Shop No. 2} \end{matrix}$$

The cost matrix, B of the different sizes of the shampoo is given by

Cost (in Rs.)

$$B = \begin{pmatrix} 40 \\ 30 \\ 20 \end{pmatrix} \begin{matrix} \text{Large} \\ \text{Medium} \\ \text{Small} \end{matrix}$$

Find the investment in shampoo by each shop.

2.16 A firm produces three sizes of bolts in two different qualities. The production (in thousands) at its Plant I is given by the following matrix:

$$\begin{matrix} & \text{Size I} & \text{Size II} & \text{Size III} \\ \text{Quality 1} & (54 & 72 & 60) \\ \text{Quality 2} & (36 & 52 & 42) \end{matrix}$$

The production (in thousands) at its Plant II is given by the following matrix:

$$\begin{matrix} & \text{Size I} & \text{Size II} & \text{Size III} \\ \text{Quality 1} & (64 & 80 & 70) \\ \text{Quality 2} & (50 & 76 & 60) \end{matrix}$$

- (i) Write a matrix that represents the total production of bolts at both plants.

- (ii) The firm's management is planning to open a third plant which would have one and one-half times the capacity of its plant I. Write the matrix representing the production at the third plant.
- (iii) What will be the production of all the three plants?

2.17 The annual sale volumes of three products X,Y,Z, whose sale prices per unit are Rs.3.50, Rs.1.50 respectively, in two different markets I and II are shown below:

Market	Product		
	X	Y	Z
I	6000	9000	13000
II	12000	6000	17000

Find the total revenue in each market with the help of matrices.

(From past C.A. examination, June 6, 1993)

2.18 A manufacturer is manufacturing two types of products A and B. L_1 and L_2 are two machines, which are used for manufacturing these two types of products. The time taken both by A and B on machines is given below:

	L_1	L_2
Product A	Product B	20 Hours
10 Hours	10 Hours	20 Hours

If 600 hours is the time available on each machine, calculate the number of units of each type manufactured using matrix method only. **(From past C.A. examination, Nov 4, 1994)**

2.19 In a certain city, there are 5 colleges and 20 schools. Each school has 3 peons, 1 clerk and 1 Head Clerk, whereas a college has 5 peons, 3 clerks, 1 Head Clerk and a additional staff as a caretaker. The monthly salary of each of them is as follows:

Peon = Rs.1100; Head Clerk = Rs.3000

Clerk = Rs.1700; Caretaker = Rs.2000

Using matrix method, find the total monthly salary bill of each School and College.

(From past C.A. examination, May 7, 1996)

2.20 Three firms A, B, C supplied 40, 35, and 25 truck loads of stones and 10, 5, 8 truck loads of sand respectively to a contractor. If the cost of stone and sand are Rs.1200 and Rs.500 per truckload respectively, find the total amount paid by the contractor to each of these firms, by using matrix method. **(From past C.A. examination, Nov 2, 1996)**

2.21 The total cost function of a firm is: $C = \frac{1}{3}x^3 - 5x^2 + 28x + 10$ where C is total cost and x is output. A tax at the rate of Rs.2 per unit of output is imposed and the producer adds it to his cost. If the market revenue function is given by: $px = (2530 - 5x)x$, where Rs.p is the price per unit of output, find the profit maximizing output and price.

2.22 A company produces x units of $\frac{1}{3}x^3 - 18x^2 + 160x$ output at a total cost of Find:

- (i) Output at which marginal cost is minimum
- (ii) Output at which average cost is equal to marginal cost.

2.23 Given the total cost function as: $C = ax^2 + bx + c$, where $a > 0$, $b^3 > 0$, $c^3 > 0$. show that average and marginal cost are equal at minimum average cost.

2.24 A firm produces 2 tonnes of output at a total cost:

$$C = \text{Rs} \left(\frac{1}{10}x^3 + 5x^2 + 10x + 5 \right)$$

At what level of output will the marginal cost and the average variable cost attain their respective minima?

2.25 If the cost function is $C(x) = 4x + 9$ and the revenue function is $R(x) = 9x - x^2$, where R and C are measured in millions of rupees, find the following:

- (i) Marginal revenue.
- (ii) Marginal revenue at $x = 5$, $x = 6$.
- (iii) Marginal cost.
- (iv) The fixed cost.
- (v) The variable cost at $x = 5$.
- (vi) The break – even point, that is $R(x) = C(x)$.
- (vii) The profit function.
- (viii) The most profitable output.
- (ix) The maximum profit.
- (x) The marginal revenue at most profitable output.
- (xi) The revenue at the most profitable output.
- (xii) The variable cost at the most profitable output.



3

Collection and Presentation of Data



Structure

- 3.1 What is Data
- 3.2 Types of Data and its Sources
- 3.3 Collection of Primary Data
 - 3.3.1 Designing a Questionnaire
- 3.4 Presentation of Data
 - 3.4.1 Ordered Array
 - 3.4.2 Pictorial Presentation of Data
 - 3.4.2.1 Diagrammatic Representation of Data
 - Bar Diagram
 - Pie Diagram
 - Pictogram
 - Population Pyramid
 - Flowchart
 - 3.4.2.2 Graphical Representation of Data
 - Graphs of time series or line graphs
 - Area Graph
 - Scatter graph
 - Graphs of frequency distribution
 - 3.4.3 Frequency Distribution
 - 3.4.4 Relative Frequency Distribution
 - 3.4.5 Cumulative Frequency
 - 3.4.6 Graphical Representation of Frequency Distribution
 - 3.4.7 Stem and Leaf Display
- 3.5 Caselet
- 3.6 Excel Guide
- 3.7 Exercises

3.1 WHAT IS DATA?

Data, the raw material of any enquiry, constitute the foundation of statistical analysis. It is the collection of facts, concepts or instructions in a formalized manner suitable for communication or processing by human or automatic means from which conclusions may be drawn. A collection of data is known as a data set and a single observation a data point. Data is said to have become information when it is organized respect to a certain context. Any statistical investigation is a comprehensive process and need systematic collection of data. By different statistical analysis, data can be changed into information that is very helpful in managerial decision-making. The quality of data greatly affects the decision making process of a fact for which statistical investigation is planned and hence proper concentration on it's accuracy is a must while collecting it.

3.2 TYPES OF DATA AND ITS SOURCES

Data can be classified into various ways depending on the sources and the nature. Some of the important categories are:

Primary and Secondary Data

Data, which are collected directly from the field of enquiry i.e., absolutely original in nature are known as *primary data*. On the other hand, data which have already been collected and recorded by others and are now being used by some one else, is known as *secondary data*. The same data is *primary* when it is in the hand of the collecting authority but are secondary in the hands of others. The simplest example is the census data of India. When it is collected, it is *primary* in the hands of the government authority and becomes *secondary* when any other secondary authority uses it. The field of investigation itself is the source of *primary data*. Journals, reports, Government and non-Government publications etc. are the sources of *secondary data*.

Discrete and Continuous Data

A set of data is said to be *discrete* if the values or observations belonging to it are distinct and separate, i.e. they can be counted or categorized. The examples of discrete data are number of students in the class (20, 30 —), gender (male, female), blood group (O, A, B, AB) etc. Discrete data may be quantitative or qualitative in nature. In the above examples the number of students is *quantitative discrete* data and the other two are *qualitative* in nature.

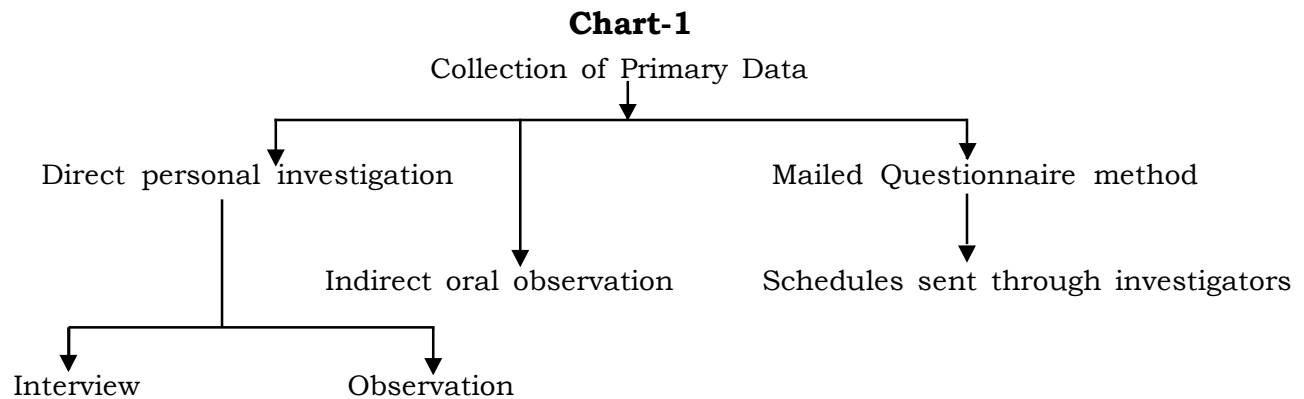
A set of data is said to be *continuous* if the values or observations belonging to it may take on any value within a finite or infinite interval. *Continuous data* can be ordered and measured. For example height, weight, temperature, rainfall etc. are all *continuous data*.

Internal data and External Data

Internal data originate within the business in the process of keeping records such as the amount of raw material used, salaries of the employee, sales amounts of the product etc. Thus, the chief sources of internal data are the records kept by the organization it self. Since internal data originates within the organization collection of it doesn't create much difficulty but the accurate recording is very important for the authenticity of this type of data. External data on the other hand are originated from outside field of enquiry. As compared to internal data it is difficult to collect.

3.3 COLLECTION OF PRIMARY DATA

Any statistical investigation follows certain stages viz. collection of data, organization and presentation of it, analysis of data and lastly the interpretation. At the very beginning there is the collection of data. The basic methods, which are commonly used for collecting primary data: - (a) Direct personal investigation, (b) Indirect oral observation, (c) Mailed questionnaire method, (d) Schedule sent through investigators.



(a) Direct personal investigation again is of two types viz. *personal interview* and *personal observation*. In *personal interview* method, the investigator personally approaches each informant and gathers the required information. Under the observation method rather than asking anybody, the investigator personally observes and records the information related to a particular field.

(b) In indirect oral observation, instead of directly approaching the actual field or person, data are collected from third party informant who are supposed to have the information about the problem under investigation.

(c) In mailed questionnaire method well-prepared questionnaires (set of questions relevant to the subject of investigation) are mailed to a selected list of persons with the request to return them duly filled in.

(d) Schedule sent through investigators is the most widely used method of collecting primary data. In this method the selected enumerators go to the respondent with a schedule (questionnaire) specially designed for the purpose, interview them and fill the schedule themselves on the spot depending on the answers received from them.

3.3.1 Designing a Questionnaire

In case of primary data collection, the *questionnaire* is the most important media of communication between investigator and respondents. The accuracy of data collected from questionnaire to a large extent depends on the designing of it. How to prepare a good questionnaire? What are the relevant precautions the enumerator should take while preparing a questionnaire? Now, we briefly discuss these questions:

Types of Questions

The following points must be kept in mind, while framing the questions.

1. The number of question should be as few as possible.
2. Questions should be of objective type as far as possible. Yes or no type or simple tick marking answers are preferred.

3. Questions should be properly arranged to have a systematic and easy flow of answer in turn.
4. Questions affecting the sentiment and pride of the respondent should be avoided.
5. To make it easy for the respondent to answer, necessary instructions and guidelines should be provided.

Types of Questionnaires

(i) Structured or Non Structured:

A structured questionnaire consists of a set of questions arranged in a predetermined order or segmented. Each question requires the respondent to make a choice among a few given predetermined responses. For example a question could be

How frequently do you go to watch a movie?

Very Frequently Often Sometimes Never

Such questions are called closed questions

A non structured questionnaire is contrast, consists of what are called open-ended questions. Examples of such questions are

How do you spend your free time?

How would you describe the ambience of the new store?

Such questions give the respondent freedom to answer according to their views and opinions.

(ii) Disguised and Non-Disguised questionnaire:

In a non-disguised questionnaire, the purpose or objectives of the study are made known to the respondent while in a disguised questionnaire, the respondents are not taken into confidence regarding the purpose or objectives of the study. A disguised questionnaire is not very popular as respondents may not be forthcoming in their answers when they do not know the objectives or relevance of the questions or the study.

3.4 PRESENTATION OF DATA

Data prior to it being arranged or organized in a proper form is called raw data.

Example of Raw Data

Table-3.1 shows the average daily earnings of a sample of 50 shopkeepers of a medium size mall. (in Rs' 00)

Table 3.1
Average Daily Earnings of 50 shopkeepers (in Rs' 00)

195	187	178	202	202	190	218	213	172	186
180	185	185	190	192	165	156	181	205	189
183	188	170	221	178	194	161	173	162	220
185	201	190	183	155	169	182	158	150	200

From this raw data of Table 3.1 it is not easy to project the trend and the other relevant characteristics about the observations. Arranged data in compact and usable form becomes the helpful devices to decision makers in making intelligent decisions. So, once the collection part is over, the next step is to present the data through some appropriate form. A variety of tables, graphs and charts are available for data presentation. The choices depend on the nature and the uses to which it is to be put. For example if a company requires a detailed analysis on the monthly sales over the last one year, then presentation by means of a table is the most effective method. If it is required to highlight the pattern of change in the sales over time, a graph is the most suitable device. Again if a company needs to portray the breakdown of sales by product or market, some form of charts should be used.

3.4.1 Ordered Array

The simplest form in which the raw data can be arranged is the ordered array. An ordered array consists of an ordered sequence of raw data in rank order from the smallest observation to the largest observation. The data of Table 3.1 is presented in the array form in Table 3.2 below:

Table 3.2

Ordered Array of the Average daily earnings (Rs.) of 50 shopkeepers

150	158	164	170	180	184	186	190	195	205
150	161	165	172	181	185	187	190	200	213
155	161	169	173	182	185	188	192	201	218
155	162	170	178	183	185	189	194	202	220
156	163	170	178	183	185	190	195	202	221

Example 3.1: The data in the following table displayed the electricity cost (00Rs.) during the month of February 2006 of 50 households of a colony. Represent the data in array form.

Table 3.3

Electricity costs of 50 households of a colony

10	5	10	40	22	26	15	17	44	43
25	7	11	49	25	27	16	20	45	25
30	22	9	39	29	28	22	22	18	29
40	25	23	40	9	32	18	32	12	16
50	12	24	42	30	42	21	41	8	17

Solution: The data in array (from lowest cost to the highest cost) is presented below:

Table 3.4
Ordered array of electricity costs of 50 households of a colony.

5	10	15	18	22	25	27	30	40	43
7	10	16	18	22	25	28	32	40	44
8	11	16	20	22	25	29	32	41	45
9	12	17	21	23	25	29	39	42	49
9	12	17	22	24	26	30	40	42	50

3.4.2 Pictorial Presentation of Data

We will discuss the pictorial representation of data by dividing it into two broad groups, Diagram and Graphs. Diagrams are more suitable to represent discrete data while continuous data are better represented by graphs. The following diagrammatic and graphic methods are most popular and commonly used for presenting data:

(1) Diagrammatic representation

- Bar diagram
- Pie diagram
- Box and Whisker diagram
- Pictogram
- Population Pyramid
- Flowchart

(2) Graphic representation

- Graphs of time series or line graphs
- Area Graph
- Scatter graph
- Graphs of frequency distribution

The different types of diagrams and the graphs are illustrated in this section itself and the graphical representations of frequency distribution are discussed in section 3.4.3 just after a discussion on frequency distributions.

3.4.2.1 Diagrammatic Representation of Data

Bar Diagram

Bar diagram consists of some equidistant rectangular bar with the lengths equal to the values of the variables. The widths of the bars are not so important, but all bars should be of equal width. The following are the different forms of bar diagram:

Simple bar diagram

Example 3.2: To represent a single variable in a simplified way this bar diagram is used.

Suppose the following were the gross revenues (in Rs' 0000) of a company XYZ for the years 2000 to 2004.

Table 3.5
Gross Revenue of XYZ company (in Rs' 0000)

Year	Revenue
2000	220
2001	300
2002	380
2003	480
2004	550

The bar diagram for this data can be drawn as follows:

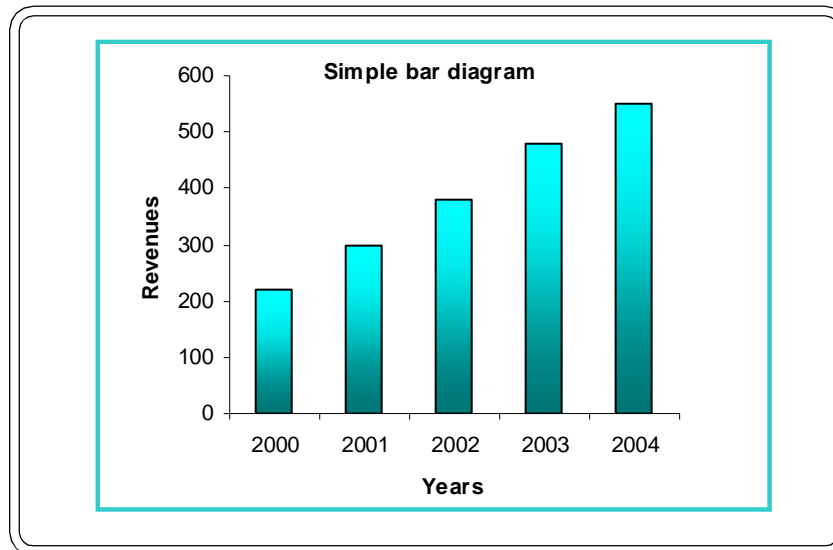


Figure-3.1

Simple Bar Diagram

Subdivided bar diagram

A *sub divided bar* diagram is used to represent various parts of a total. In this diagram the simple bars are further sub divided into different components based the given information.

Example 3.3: The following table shows the percentage outlay of first two five year plans. Represent the data by means of a sub-divided bar chart.

Table 3.6
Percentage Outlay of first two five year plan

Items	Outlay (percentage)	
	Plan I	Plan II
Agriculture	14.8	11.7
Irrigation	29.7	18.9
Industry	5.0	24.1
Transport	26.4	27
Social service	21	16.5
Miscellaneous	3.1	1.8
Total	100	100

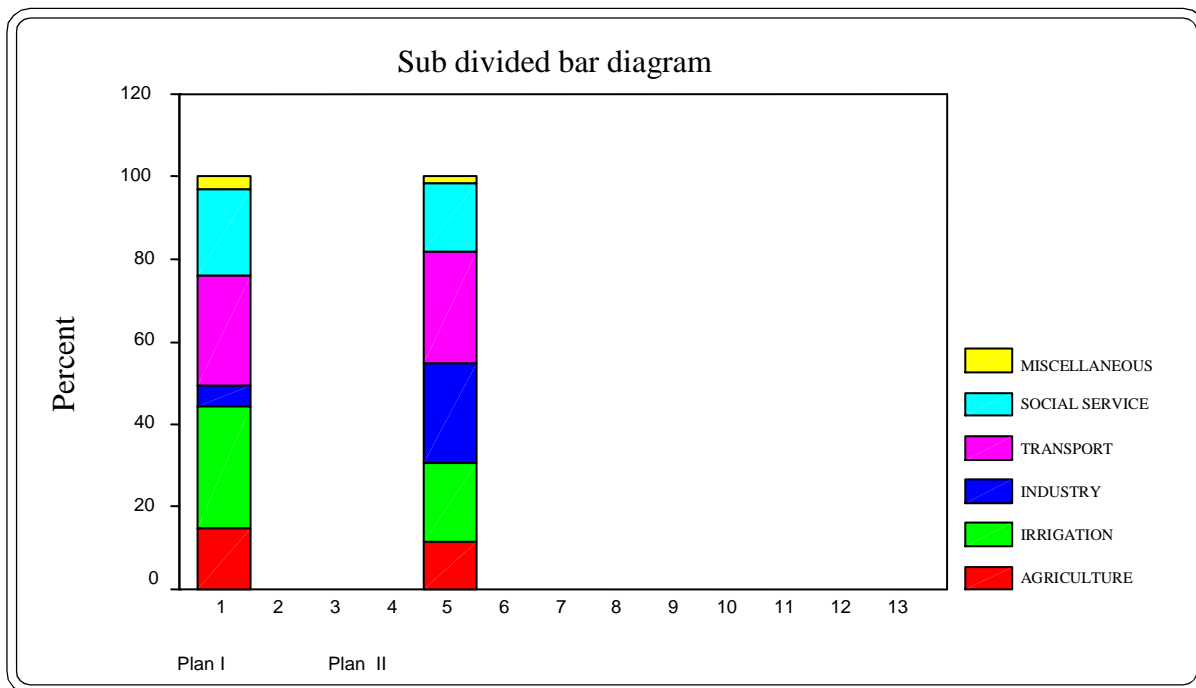


Figure-3.2

Sub Divided Bar Diagram

Multiple bar diagram

More than one set of interrelated data is represented by the multiple bar diagram.

Example 3.4: Represent the following data by a suitable diagram showing the picture of proceeds and costs of a firm for the last five years:

Table 3.7
Proceeds and costs of a firm

Year	Proceeds Cost (Thousand rupees)	
2000	72	69.5
2001	77.3	71.7
2002	78.2	80
2003	80.3	75.6
2004	82.3	76.1

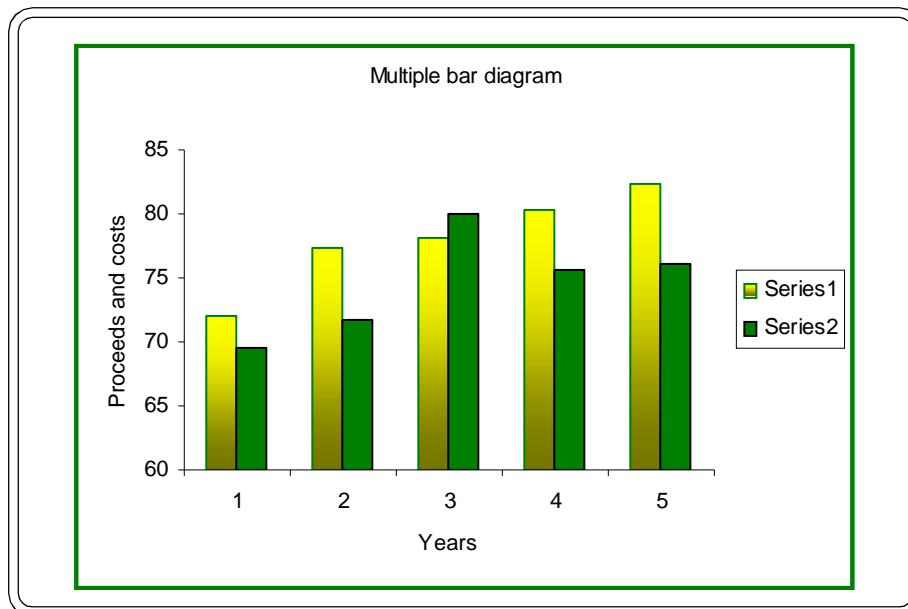


Figure 3.3

Multiple Bar Diagram

Pie Diagram

Pie diagram is in the form of a circle or a *pie* whose area is divided proportionately among the different components of a variable. The pie chart is based on the fact that the circle has 360° . The total area of the pie is divided into different parts according to the percentage in each item.

Example 3.5: The following table shows the areas of the continents in the World

Table 3.8
Areas of various continents

Continent	Area (Million square kilometer)	Percentage area	Share in the pie (Degree)
Asia	26.9	20.18	72.65
Africa	30.3	22.73	81.83
Europe	4.9	3.68	13.23
North America	24.3	18.23	65.63
South America	17.9	13.43	48.34
Russia	20.5	15.38	55.36
Oceania	8.5	6.37	22.96

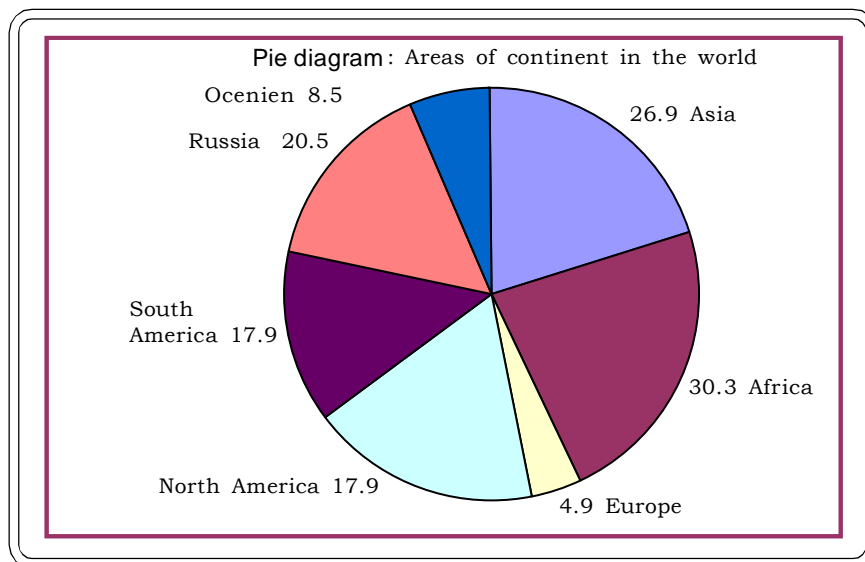


Figure-3.4

Pie Diagram

Thus, to construct a pie diagram, the individual area has to be expressed in terms of 360° circular arc. For example, for Asia the share of arc will be $(360^\circ/133.3) \times 26.9 = 72.65^\circ$.

Example 3.6: The following table gives the construction cost of a shop. The components are given below. Represent the data by Pie-diagram.

Table 3.9.
Construction cost of a shop

Category	Cost ('000 Rs.)
Cement	543
Bricks	470
Electricity	286
Plumbeing	220
Miscellaneous	100

Solution: To represent this information in the form of a pie-chart, we first calculate the degree equivalent of each category in a 360° circle as follows:

Category	Cost (Rs' 000)	Percentage	Share in the pie (Degree)
Cement	543	$33.5\% \cong 34$	$34 \times 3.6 = 122.4$
Bricks	470	$29\% \cong 29$	$29 \times 3.6 = 104.4$
Electricity	286	$17.66\% \cong 18$	$18 \times 3.6 = 64.8$
Plumbing	220	$13.5\% \cong 13$	$13 \times 3.6 = 46.8$
Miscellaneous	100	$6.17\% \cong 6$	$6 \times 3.6 = 21.6$
	1619		

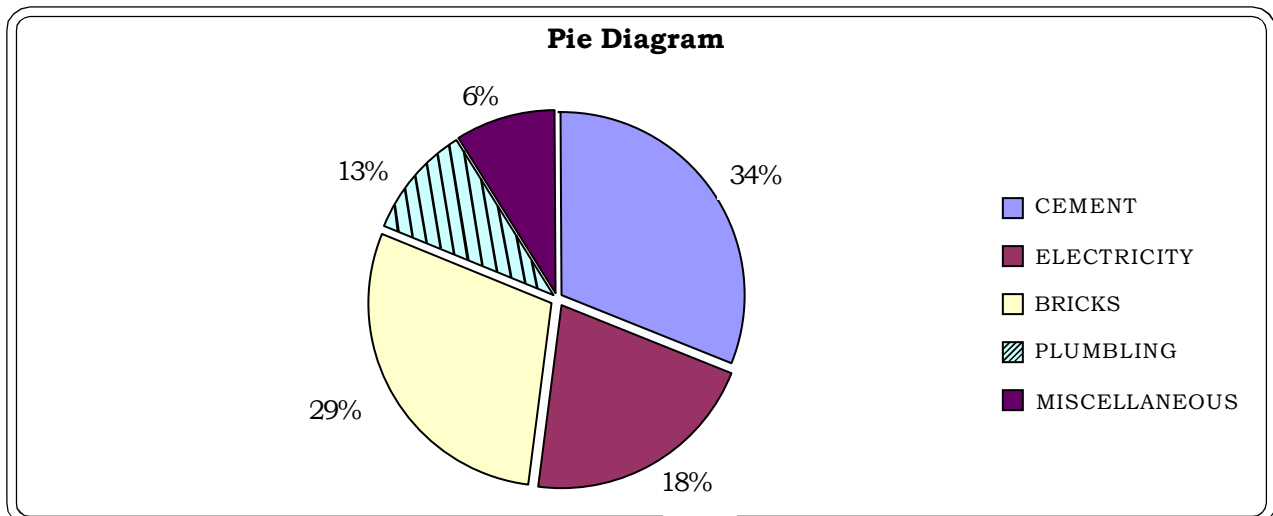


Figure 3.5

Pie Diagram

Pictogram

In a *Pictogram*, the frequencies are indicated by a number of identical pictures. While using a pictogram, it is a must to specify what the individual pictures represent.

Example 3.7: The following table shows the sale of ice creams of different flavour in a particular shop.

Table 3.10
Sales of the Creams

Flavour	Number of ice cream
Vanilla	10
Strawberry	9
Raspberry	5
Others	7

Solution:

The resulting pictogram is as follows.

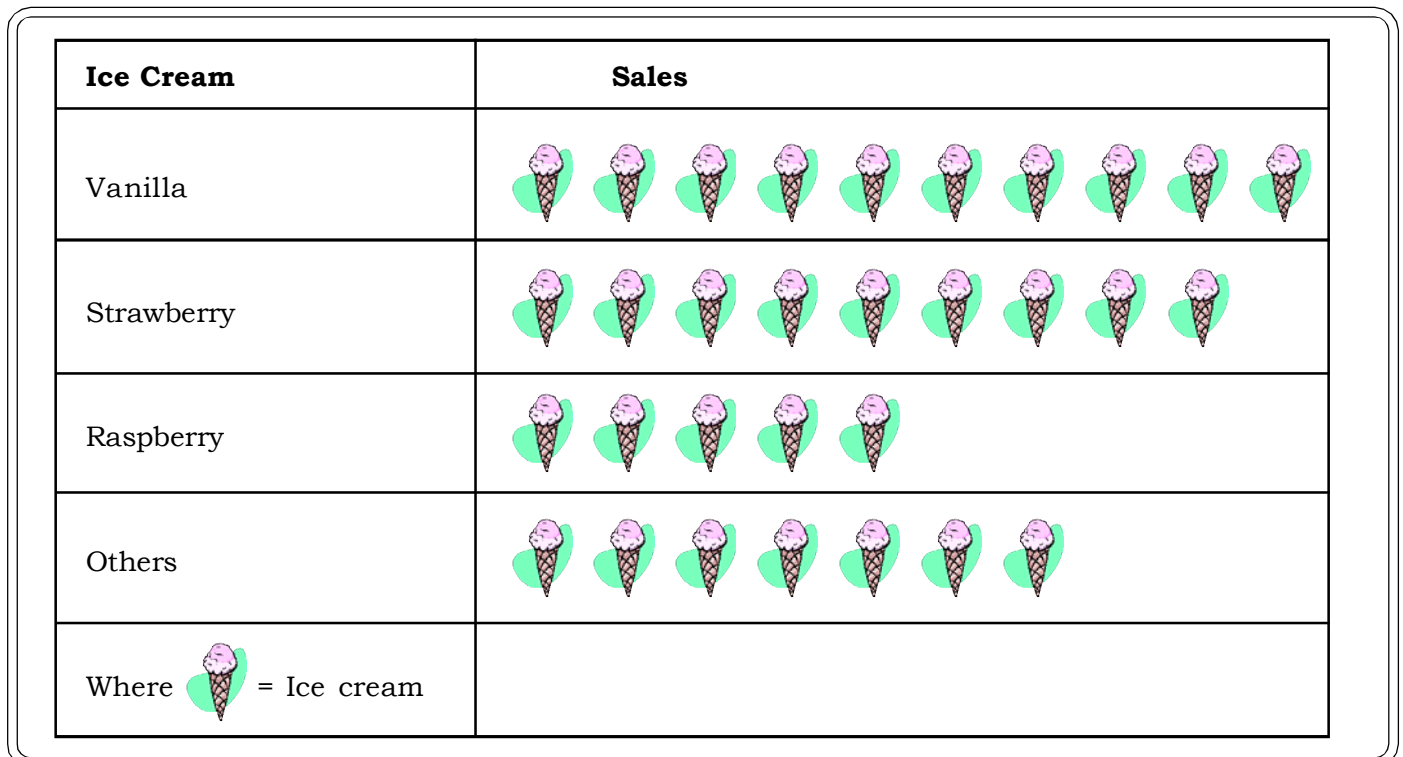


Figure 3.6

A Pictogram

Population Pyramid

The age-gender distribution of a population is an important feature to analyse if we wish to understand a country's demographic situation. These statistics give governments and others one of the tools they need to make informed decisions that will affect our lives today and in the future. A handy way to illustrate the structure of a population is to graph the number of males and females for various ages. A horizontal bar graph with data for males on the left and females on the right is called a population pyramid.

Example 3.8: Create a population pyramid from the following data

Table 3.11
Percentage Distribution of males and females in different age groups.

Age Group	Male %	Female%
0-9	12	8
10-19	9	6
20-29	7	5
30-39	13	12
40-49	9	7
50-59	7	5

Solution:

The population pyramid depicting the distribution of males and females in different age groups is constructed below:

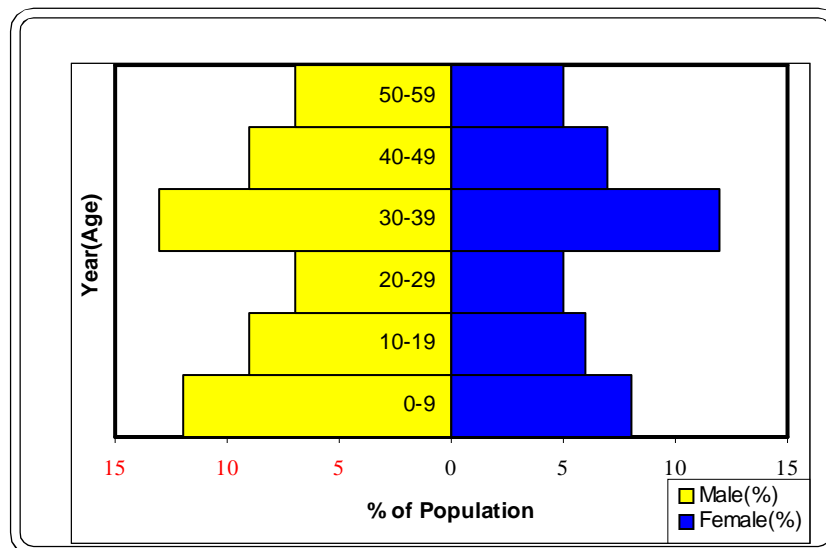


Figure-3.7

A Population Pyramid

Flow Charts

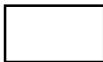
Flow charts are used most commonly to represent the internal logical organization of computer programs or in a manufacturing firm to organize the hierarchy of a production process. However, they can be used in any situation where we wish to represent connected structures where there may be alternative pathways through the system. We can also indicate quantitative aspects of the flow of information or materials through the structure by annotating the diagram, or varying the line style or thickness to indicate quantity.

Standard symbols for flowchart

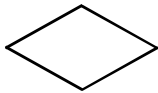
It is not strictly necessary to use boxes, circles, diamonds or other such symbols to construct a flowchart, but these do help to describe the types of events in the chart more clearly. Described below are sets of standard symbols, which are applicable to most situations without being overly complex.



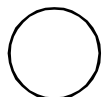
Rounded box - used to represent an event, which occurs automatically. Such an event will trigger a subsequent action, for example 'receive telephone call', or describe a new state of affairs.



Rectangle or box - used to represent an event, which is controlled within the process. Typically this will be a step or action, which is taken. In most flowcharts this will be the most frequently used symbol.



Diamond - used to represent a decision point in the process. Typically, the statement in the symbol will require a 'yes' or 'no' response and branch to different parts of the flowchart accordingly.

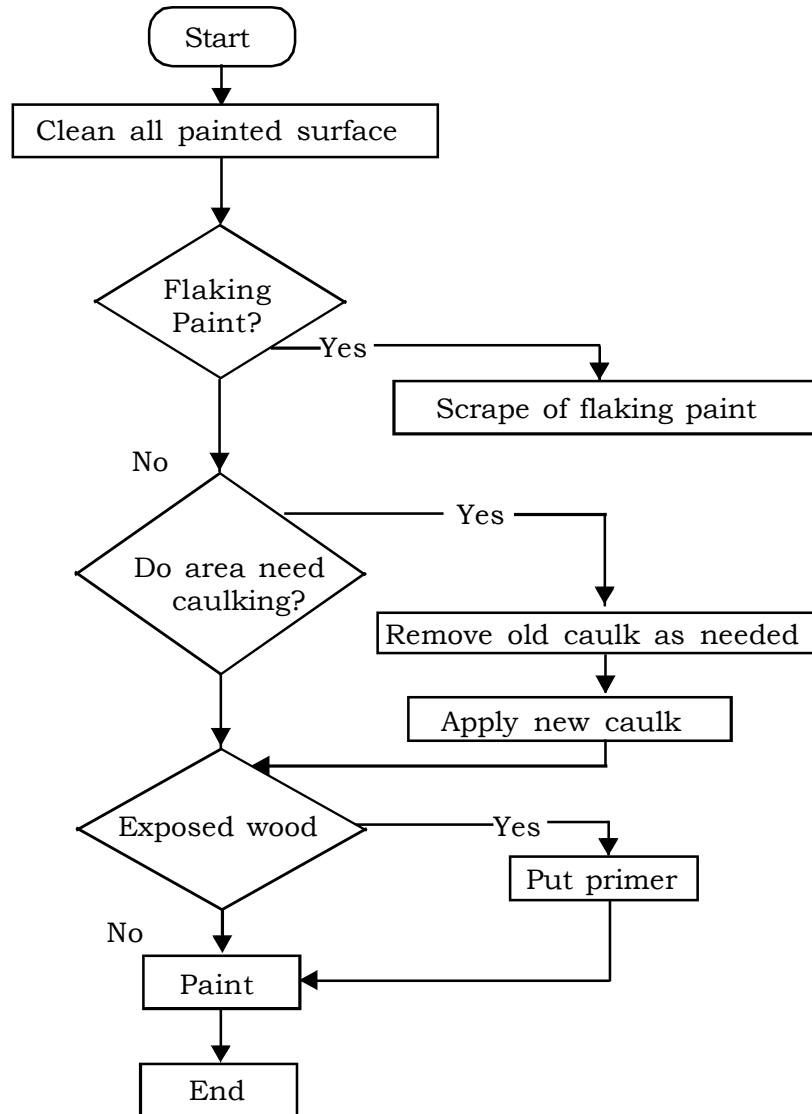


Circle - used to represent a point at which the flowchart connects with another process. The name or reference for the other process should appear within the symbol.

The example of a flow chart shown in Chart 2 uses these symbols to indicate the flow-charting of a house painting, process.

Chart-2

Flow Chart of a house painting process



3.4.2.2 Graphical Representation of Data

Graphs of time series or line graphs

This is one of the very common and popular methods of representing statistical data, especially used in business and commerce where data are shown in accordance with the time of occurrence. The line graph shows by means of a curve or straight line, relationship between two variables over time. Suppose we want to plot the expenditure and revenue of a company from 1998 to 2005 in order to identify the trend of profit change during the time period. The table below provides the required data. These data are graphically represented by line graph in Figure-3.8.

Table 3.12
Expenditure and Revenue of a company

Year	Expenditure (Rs. Lakhs)	Revenue (Rs. Lakhs)
1998	10	15
1999	14	18
2000	13	19
2001	15	20
2002	18	21
2003	20	26
2004	21	38
2005	25	40

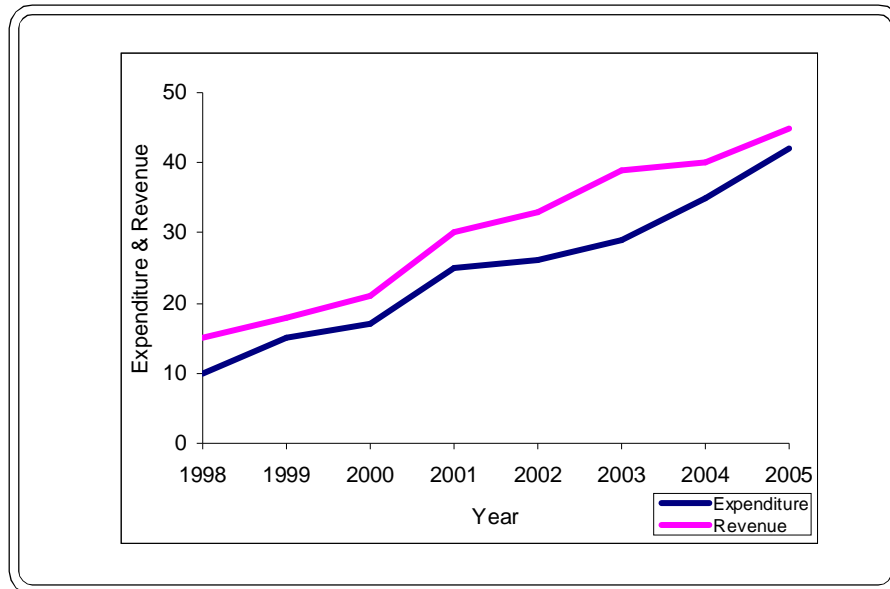


Figure-3.8

Line Graph

Area Graph

Area charts (as in Figure 3.9) are similar to line charts, but are used for continuous data where there is one (continuous) independent series, and several dependent series. The latter together have a *constant sum*, such as the region wise exports of a country, as given in the table below:

Table 3.13
Regionwise exports of a country

Amount (Lakhs, Rs)

Export Item	2000	2001	2002	2003	2004
Region-1	25	22	18	28	30
Region-2	40	41	51	43	32
Region-3	10	11	12	20	15
Region-4	4	5	3	9	5
Region-5	79	79	84	100	82

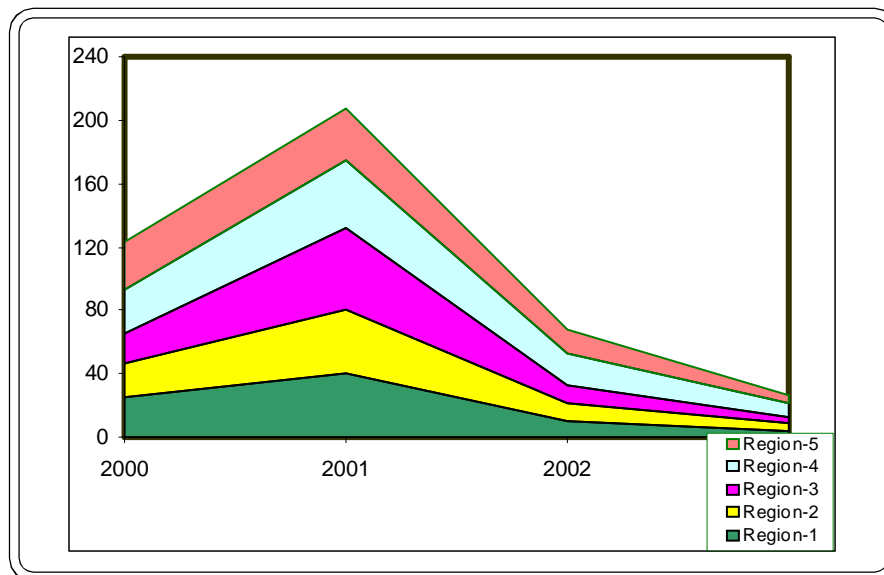


Figure-3.9

Area Graph

Scatter graphs

Scatter graphs are widely used in science to present measurements on two variables that are thought to be related. In a scatter graph the values of the *dependent* variables are measured on the vertical axis and the values of the *independent variable* plotted along the horizontal axis.

The origin of the graph - the point at which the axes cross - should *almost always* be (0,0). Figure 3.11 is an example of a scatter graph.

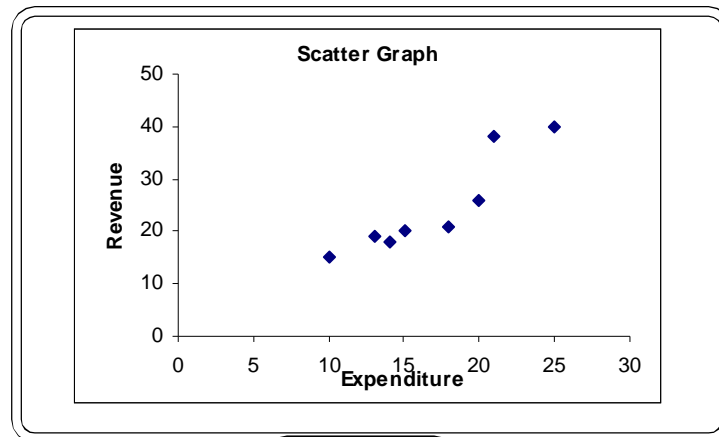


Figure-3.10

Scatter Graph

Example 3.9: The ten years production data of a company is given below. Draw a scatter plot on it.

Table 3.14
Ten Year Production Data.

Years	Production (lakhs Rs.)
1997	10
1998	12
1999	15
2000	16
2001	18
2002	17
2003	22
2004	21
2005	30

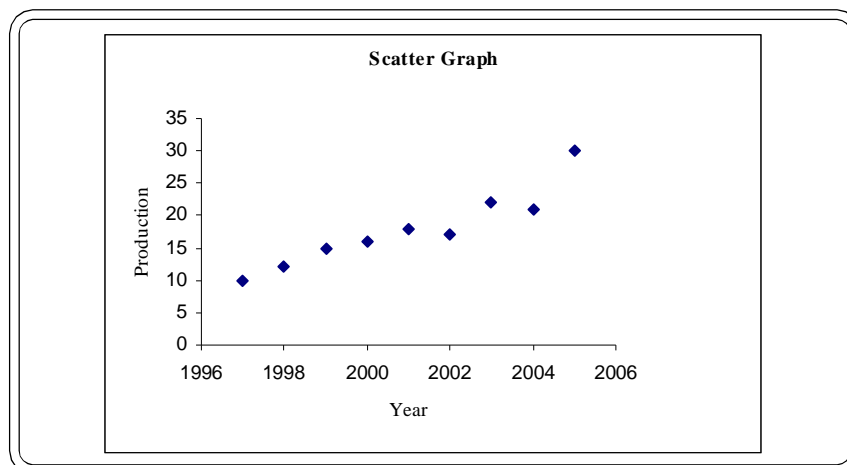


Figure 3.11

Scatter Graph

3.4.3 Frequency Distribution

One way to represent a large number of data in a summarized form is the frequency distribution. Frequency distribution is nothing but a statistical table that shows the values of the variable arranged in order of magnitude either individually or in groups and the corresponding frequencies side by side. In case of individual presentation or *simple frequency distribution* (Table-3.15) the values of the variable are shown individually and in case of *grouped frequency distribution* (Table-3.17) data are organized into appropriate number of mutually exclusive, non-overlapping classes.

Table 3.15
Simple Frequency Distribution

Age of children in a nursery class	Frequency (Number)
2	5
3	10
4	15
5	4
6	2
Total	36

Important terms related to grouped frequency distribution

Class interval: Each of the group in a frequency distribution is defined by an interval called Class interval.

In Table 3.17 column (1) represents the class intervals as 150-160, 160-170-, ———.

Class limit: The class intervals are defined by two extreme limits known as class limits. The lower one is the lower class limit and the upper one is the upper class limit. In Table-3.17, 150 and 160 respectively are the lower and upper limits of the first class.

Class boundaries: The most extremes values, which are not covered by the class intervals, are called *class boundaries*.

$$\text{Lower class boundary} = \text{Lower class limit} - 1/2d$$

$$\text{Upper class boundary} = \text{upper class limit} + 1/2d$$

Class frequency and total frequency: The number of observations falling within a class is called its *class frequency*. *Total frequency* is the sum of all class frequencies

$$\text{Mid value or Class mark} = (\text{Lower class limit} + \text{Upper Classes limit})/2$$

$$= (\text{Lower Class boundaries} + \text{Upper Class boundaries})/2$$

Width = Upper class boundary-Lower class boundary

Frequency density = Class Frequency/Width

Steps of constructing a grouped frequency distribution

Step-1: Find the largest and the smallest observation of the given data set.

Step-2: Calculate the Range, the difference between the largest and the smallest observation.

Step-3: Divide the Range into a suitable self-selected number of class intervals. Normally the number of classes should lie in between 5 and 15 depending on the numbers of the observations. The class limits are so chosen that the lowest observation will lie in the first cell and the highest in the last. Preferably the classes will be of equal width.

Step-4: Use *tally mark* to facilitate the counting of the number of observations falling in a particular class.

Step-5: Now prepare the final frequency distribution table showing class intervals or class boundaries in the first column and the corresponding frequencies in the second.

(The method of constructing frequency distribution through excel is presented in section 3.6).

Example 3.10: Consider the raw data of Table 3.1 giving the average daily earnings of 50 shopkeepers and represent it into a suitable frequency distribution

Solution:

Maximum value = 221

Minimum value =150

Range = 221 -150 = 71

Let start with the class limits 150-160, 160-170, each of width 10. Now use tally marks to tabulate the class frequencies. The tally marks are shown in a group of five. Every fifth tally will be placed across the previous four.

Table 3.16
Frequency Distribution Table

Class Limits	Tally marks	Frequency
150-160	/	6
160-170		7
170-180		7
180-190		14
190-200		7
200-210		5
210-220		2
220-230		2
Total		50

The final frequency distribution is as follows:

Table 3.17
Grouped Frequency Distribution

Earnings (Rs)	Frequency
150-160	6
160-170	7
170-180	7
180-190	14
190-200	7
200-210	5
210-220	2
220-230	2
Total	50

Example 3.11: Form a frequency distribution for the following data of age of 40 employees of a Delhi based private sector organization.

Table 3.18
Age of Employees of an organization

50	58	41	56	55	24	40	32
30	59	43	54	29	26	30	27
22	32	42	28	49	27	35	33
29	31	44	27	33	39	45	36
40	40	51	30	29	44	44	42

Solution:

The Maximum Value = 58

The Minimum Value = 22

Range = $58 - 22 = 36$

Let we start with the class 20 – 25 i.e. starting from 20 with class intervals of size 5.

Calculation of Tally marks

Table 3.19
Frequency Distribution Table

Class Limits	Tally marks	Frequency
20-25	//	2
25-30	 	8
30-35	 	8
35-40		3
40-45	 	10
45-50	//	2
50-55		3
55-60		4
Total		40

We form the frequency table:

Table 3.20
Grouped Frequency Distribution Table

Age	Frequency
20-25	2
25-30	8
30-35	8
35-40	3
40-45	10
45-50	2
50-55	3
55-60	4
Total	40

Example 3.12: In the following table the lengths of 40 leaves are recorded to the nearest millimeter. Construct a frequency distribution.

Table 3.21
Lengths of 40 Leaves

138	164	150	132	144	125	149	157
146	158	140	147	136	148	152	144
168	126	138	175	163	119	154	165
146	173	142	147	135	153	140	135
161	145	135	142	150	156	145	128

Solution:

Maximum value = 175

Minimum value = 119

Range = 175-119 = 56

Let us assume the class interval is of width 5. It can be arranged in the form of a frequency distribution as follows:

Table 3.22
Construction of Frequency Distribution

Class boundaries	Tally marks	Frequency
118-123	/	1
123-128	//	2
128-133	//	2
133-138	////	4
138-142	//// /	6
142-148	//// ///	8
148-153	////	5
153-158	////	4
158-163	//	2
163-168	///	3
168-173	/	1
173-178	//	2
Total		40

Final frequency table:

Table 3.23
Frequency Distribution of Length of Leaves

Length (mm)	Frequency
118-123	1
123-128	2
128-133	2
133-138	4
138-142	6
142-148	8
148-153	5
153-158	4
158-163	2
163-168	3
168-173	1
173-178	2

Up to this point of frequency distribution we have described some quantitative attribute of the items sampled. The classification can also be made in terms of qualitative characteristics like religion, qualification, occupation etc. Table 3.24 shows how to construct a frequency distribution table using qualitative data.

Table 3.24
Frequency Distribution of Qualitative Data

Occupational Class	Frequencies
Doctor	10
Engineer	20
Professor	17
Businessmen	27
Lawyer	6
Bankers	18
Others	12
Total	100

3.4.4 Relative Frequency Distribution

For enhancing the analysis of frequency distribution either the *relative frequency distribution* or the *percentage frequency distribution* can be developed.

Relative frequency is formed by dividing the frequencies in each class of the frequency distribution by the total number of observation (Table-3.25).

Then percentage distribution can be formed by multiplying each relative frequency by 100.

Table 3.25
Relative Frequency Distribution Table

Class Interval	Class Frequency	Relative Frequency	Percentage Frequency
10-14	5	.10	10
15-19	7	.05	15
20-14	12	.25	25
25-29	15	.31	31
30-34	6	.13	13
35-39	3	.06	6
Total	48	1	100

3.4.5 Cumulative Frequency

Cumulative frequency corresponding to a specified value of the variable is defined as the number of observations smaller than or greater than that value. The number of observation upto a given value is called *less than cumulative frequency* and the number of observations greater than a value is called the *more than cumulative frequency*.

A table presenting such frequencies side by side with the specified limit of the observation is called cumulative frequency distribution. This frequency distribution enables us to find out how many observations lie above or below certain values.

Example 3.13: Consider the distribution given in Table 3.17. The Cumulative frequency distribution of this problem will be as follows:

Table 3.26
Cumulative Frequency Distribution Table of Earnings (in Rs.)

Earnings (Rs)	Less than Cumulative Frequency
Less than 150	0
Less than 160	$(0 + 6) = 6$
Less than 170	$(0 + 6 + 7) = 13$
Less than 180	$(0 + 6 + 7 + 7) = 20$
Less than 190	$(0 + 6 + 7 + 7 + 14) = 34$
Less than 200	$(0 + 6 + 7 + 7 + 14 + 7) = 41$
Less than 210	$(0 + 6 + 7 + 7 + 14 + 7 + 5) = 46$
Less than 220	$(0 + 6 + 7 + 7 + 14 + 7 + 5 + 2) = 48$
Less than 230	$(0 + 6 + 7 + 7 + 14 + 7 + 5 + 2 + 2) = 50$

Earnings (Rs)	More than Cumulative Frequency
More than 150	50
More than 160	$(50 - 6) = 44$
More than 170	$(50 - 6 - 7) = 37$
More than 180	$(50 - 6 - 7 - 7) = 30$
More than 190	$(50 - 6 - 7 - 7 - 14) = 16$
More than 200	$(50 - 6 - 7 - 7 - 14 - 7) = 9$
More than 210	$(50 - 6 - 7 - 7 - 14 - 7 - 5) = 4$
More than 220	$(50 - 6 - 7 - 7 - 14 - 7 - 5 - 2) = 2$
More than 230	$(50 - 6 - 7 - 7 - 14 - 7 - 5 - 2 - 2) = 0$

Cumulative frequencies expressed in term of percentage is known as relative cumulative frequency or percent cumulative frequency and are shown in the following table:

Table 3.27
Present Cumulative Frequency of Earnings

Earnings (Rs)	Cumulative percentages	
	Less than	More than
150	0	100
160	12	88
170	26	74
180	40	60
190	68	32
200	82	18
210	92	8
220	96	4
230	100	0

Example 3.14: Consider the following example and calculate the number of cases between 115 and 135.

Class limit	90-100	100-110	110-120	120-130	130-140	140-150	150-160
Frequency	15	21	44	62	50	25	9

Solution:

The solution table can be obtained by applying simple interpolation in the cumulative frequency distribution table.

Number of cases between 115 and 135 = number of cases less than 135 – number of cases less than 115

= less than cumulative frequency corresponding to 135 – less than cumulative frequency corresponding to 115

Class boundary	Cumulative frequency (less than)
90	0
100	15
110	36
115 →	← C
120	80
130	142
135 →	← U
140	192
150	217
160	226

Following the *simple interpolation* formula we can write:

$$\frac{115-110}{120-110} = \frac{X-36}{80-36}$$

$$\text{or } \frac{5}{10} = \frac{X-36}{44}$$

$$\text{or } C = 57$$

Similarly, for less than c.f. corresponding to 135:

$$\frac{135-130}{140-130} = \frac{Y-142}{192-142}$$

$$\text{or } \frac{5}{10} = \frac{Y-142}{50}$$

$$\text{or } U = 167$$

Therefore number of cases between 115 and 135 = $167-57 = 110$

Example 3.15: From the following distribution of age group of employee of a Public Sector Organization, find out the following:

- (i) Number of employees whose age group is between 36 to 40
- (ii) Number of employees below 32 years

Table 3.28
Age Distribution of Employess of a Public sector Organisation







Age	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45	45 – 50	50 – 55
Frequency	2	7	15	20	8	4	3

Solution:

Just like the earlier example, this example can also be solved by applying simple interpolation in the cumulative frequency distribution.

No. of employee between 36 and 40 = Number of employee less than 40- number of cases less than 36

= less than cumulative frequency corresponding to 40- less than cumulative frequency corresponding to 36

Class boundary	Cumulative frequency (less than)
20	0
25	2
30	9
32 	 Y
35	24
36 	 X
40	44
42 	 Z
45	52
50	56
55	59

(i) Cumulative frequency corresponding to the age group of 40 is 44.

Now by applying simple interpolation formula the cumulative frequency corresponding to 36 can be calculated as follows:

$$\frac{36 - 35}{40 - 35} = \frac{X - 24}{44 - 24}$$

$$\text{or } \frac{1}{5} = \frac{X - 24}{20}$$

$$\text{or } 5X - 120 = 20$$

$$\text{or } 5X = 140$$

$$X = \frac{140}{5} = 28$$

Similarly, c.f. corresponding to 40 = 44

The no. of employee whose age group are in between 36 and 40 is

$$44 - 28 = 16$$

(ii) Number of employee less than 32 = less than cumulative frequency corresponding to 32

Using simple Interpolation formula:

$$\frac{32 - 30}{35 - 30} = \frac{Y - 9}{24 - 9}$$

$$\text{or } \frac{2}{5} = \frac{Y - 9}{15}$$

$$\text{or } 5Y - 45 = 30$$

$$\text{or } Y = \frac{75}{5} = 15$$

Thus, number of employees below 32 years = 15.

3.4.6 Graphical Representation of Frequency Distribution

The pictorial representations of frequency distribution are more appealing than its tabulated counterpart. It represents the data in two-dimensional picture, the horizontal axis being used to represent the characteristics and the vertical axis for the frequencies. The most popular graphical forms of frequency distributions are

1. Histogram
2. Frequency Polygon
3. Frequency Curve
4. Ogive or Cumulative Frequency Curve

1. Histogram

Histogram is a series of rectangles with the bases equal to the width of the classes. Normally it is drawn for the grouped frequency distribution. For each group one rectangle is constructed. If the classes are of equal width, the height of the rectangles erected from each width will correspond to the class frequency. For unequal width classes, corresponding frequency densities are the heights of the rectangles.

Example 3.16: The table below shows a frequency distribution of the monthly wages in (Rs'000) of 67 employees at a particular company. Draw a *histogram*.

Table 3.29

Frequency Distribution of monthly wages of 67 employees of a company

Wages (Rs'000)	Number of Employees
30-40	2
40-50	5
50-60	8
60-70	12
70-80	18
80-90	15
90-100	4
100-110	3
Total	67

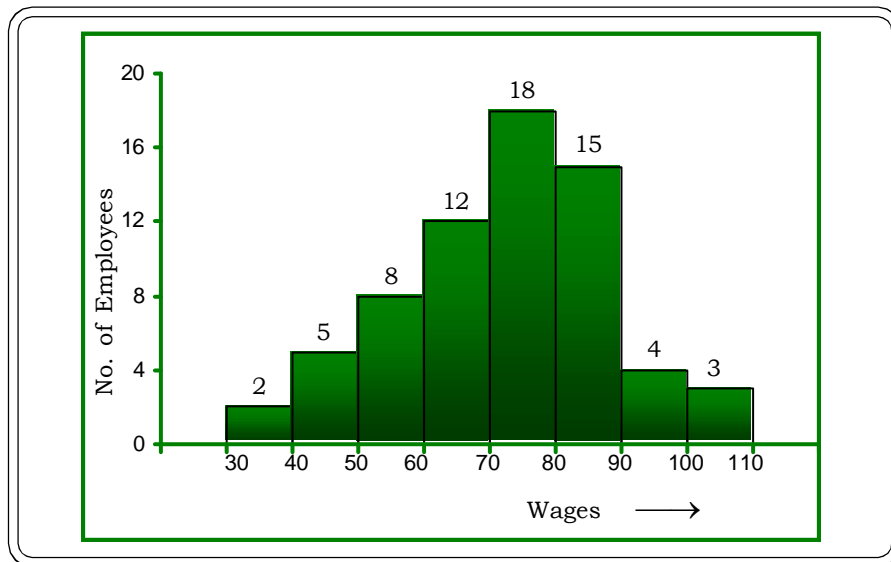


Figure-3.12

A Histogram

The histogram corresponding to the above frequency distribution of monthly wages is shown in Figure-3.12. The rectangular bars are constructed for successive class intervals with their base on the X-axis, the base being equal in width and the height on the Y-axis equal to the corresponding class frequency.

The procedure for drawing histogram in the case of unequal class intervals is slightly different from that of equal one. In such cases, the frequency densities are plotted on the Y-axis against the given class intervals. A histogram that uses relative frequencies as the heights of the rectangles is known as relative frequency histogram, which also has the same shape as a absolute frequency histogram. In case of open-ended distribution, a histogram can be drawn by dropping the open-ended class.

Example 3.17: Draw a histogram from the following frequency distribution of workers in an organization.

Table 3.30
Frequency Distribution of workers according to age

Age group (Year)	Frequency
20-25	2
25-30	3
30-35	5
35-40	9
40-45	15
45-50	8
50-55	2
55-60	1
Total	40

Solution:

Table 3.31
Frequency Distribution

Age group (Year)	Mid Values (X)	Frequency
20 - 25	22.5	2
25 - 30	27.5	3
30 - 35	32.5	5
35 - 40	37.5	9
40- 45	42.5	15
45 - 50	47.5	8
50 - 55	52.5	2
55 - 60	57.5	1
Total		40

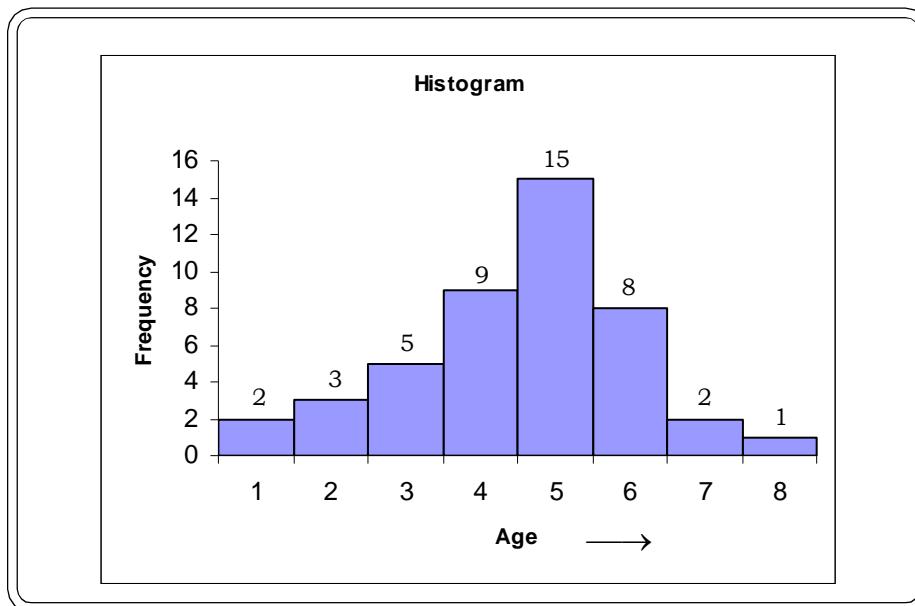


Figure 3.13

A Histogram

2. Frequency Polygon

Frequency polygon is a line graph of class frequencies plotted against class marks or mid values. It can also be obtained by connecting mid points of the tops of the rectangles in the histogram. The frequency polygon corresponding to the frequency distribution of Table-3.29 in example 3.15 is shown in Figure-3.14. Frequency polygon will touch the horizontal axis from both sides irrespective of the nature of the given observation. So it is customary to add the extensions of F and P to the next lower and higher class marks respectively.

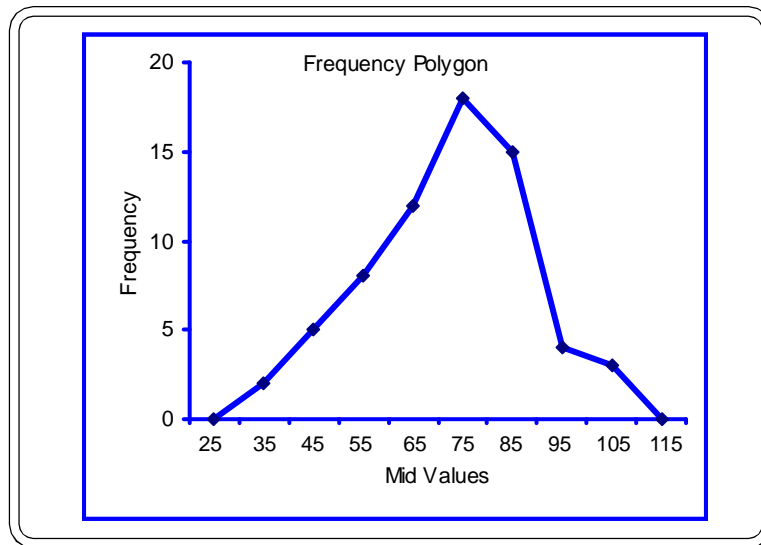


Figure-3.14

A Frequency Polygon**3. Smooth Frequency Polygon or Frequency Curves**

Smoothing a frequency polygon indicates drawing a free hand smooth curve through the points by joining which the frequency polygon is obtained [Figure-3.15]. This is done to remove the irregularities in the polygon, which may occur due to joining the various points by means of straight lines.

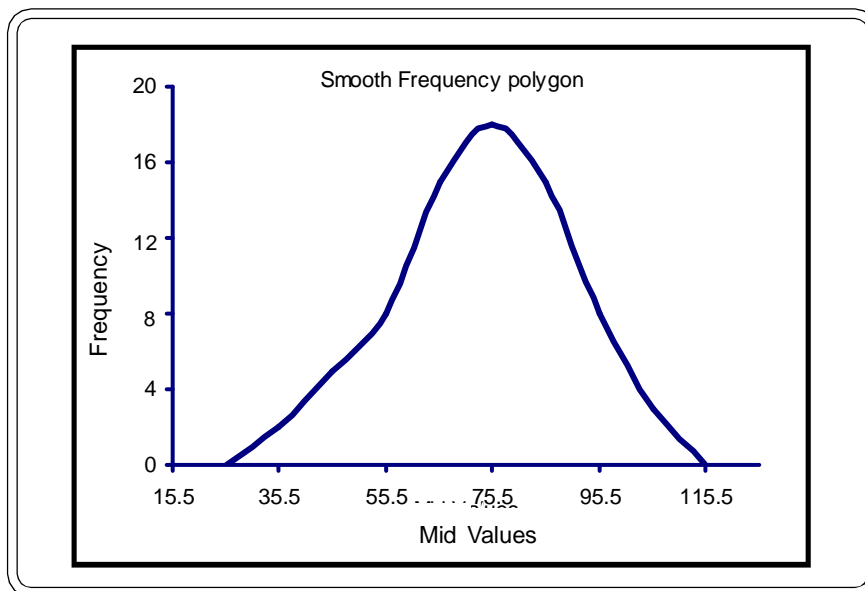


Figure-3.15

A Frequency Curve

This curve can be of various forms as indicated in the following figure-3.16

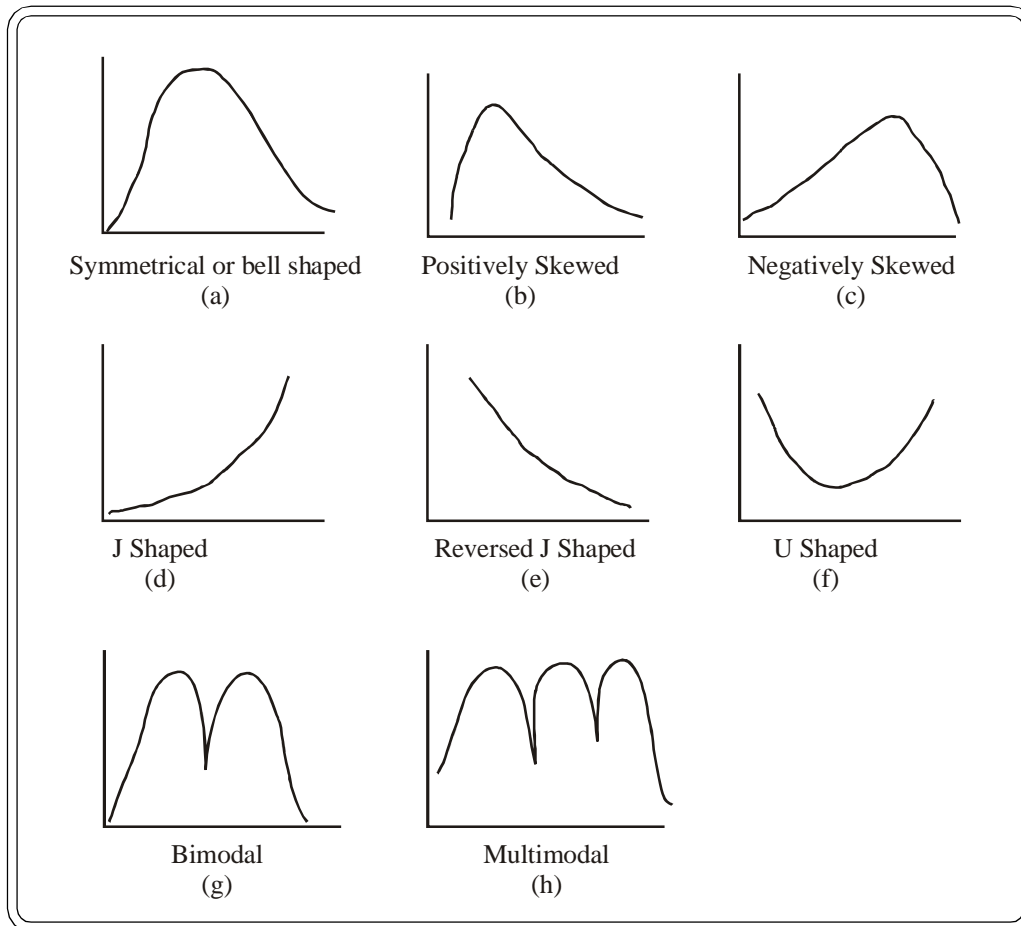


Figure-3.16

Different Frequency Curves

The *symmetrical or bell shaped* frequency curves are characterized by the fact that the observations are equidistant from the central maximum. If the longer tail occurs to the right side of the curve, the curve is *positively skewed* and for the reverse it is *negatively skewed*. In case of *J shaped* frequency curve the maximum occurs at one end. *U shaped* frequency curve has maximum at both the ends. A *bimodal* frequency curve has two maxima and the *multimodal* curve has more than two maxima points.

4. Ogive (or Cumulative Frequency Polygon)

Ogive is the graphical representation of a cumulative frequency distribution, and hence also known as Cumulative Frequency Polygon. Ogives based on the data in table 3.10 are shown in Figure-3.17.

More than ogive is plotted by taking the more the cumulative frequencies along the y-co-ordinate and the lower limit of the class interval along the x-co-ordinate.

Table 3.32
Less than and More than Cumulative Frequencies

Wages (Rs.)	Number of Employees	Less than Cumulative Frequency	More than Cumulative Frequency
30-40	2	2	67
40-50	5	7	62
50-60	8	15	54
60-70	12	27	42
70-80	18	45	24
80-90	15	60	9
90-100	4	64	5
100-110	3	67	2
Total	67		

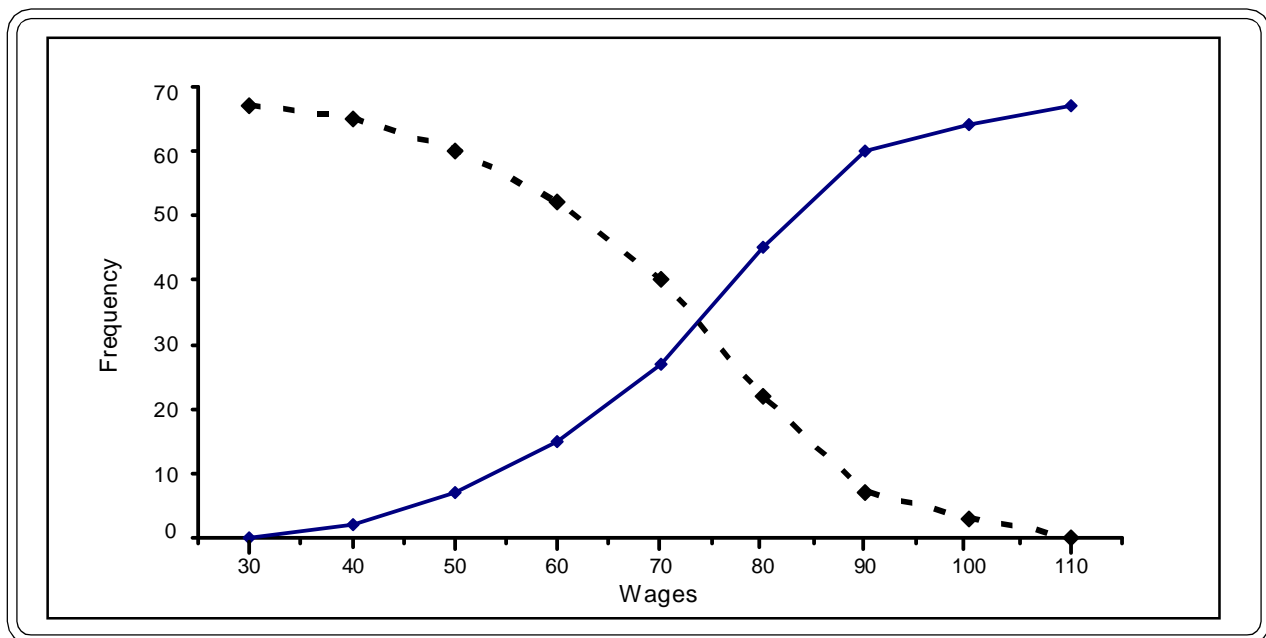


Figure-3.17

Ogive

Less than ogive is constructed by plotting the cumulative frequencies against upper limits. The plotted points are then joined by a smooth curve. The less than Ogive looks like elongated English letter 'S' and the more than type is just the reverse of that. The intersection of the two ogives gives the median of the distribution.

3.4.7 Stem and Leaf Display

This is a new and a very useful tool to study how observations of a data set distribute and cluster over the entire range. In a frequency table we do loose individual values of the observations.

Stem and Leaf display condense the data yet allows us to recover the original data set if required. The idea of this data presentation is based on the analogy of a plant. To make a stem and leaf display, the digits of each individual observation are partitioned into two components: stem and leaf. The left side group of digits of the entry is called stem and the right side group of the digits is called leaf. Both the items are presented in a table column-wise.

For example, stem may represent the tens place and leaf may represent the units place.

Example 3.18: For example to construct a stem and leaf for the following data of mid term exam scores viz. 88, 93, 56, 68, 74, 52.83, 85, 77, 79, 72, 69, 80

Table 3.33

Stem and Leaf Display of Mid-Term Examination Marks

Stem	Leaves	Frequency
5	6	1
6	8 9	2
7	2 4 7 9	4
8	0 3 5 8	4
9	3	3

From this stem and leaf display we know that the frequency in stem 7 for example is 4 and the values of the four observations are 72, 74, 77 & 79.

3.5 CASELET

The Retail industry in India has grown by leaps and bounds over the last five years. Retail is India's largest industries accounting for 10% of the country's GDP and around 8% of employment. According to experts retail consolidation is all set to be a powerful catalyst in the overall development of the Indian economy. The major players in the retail sectors in India are Pantaloon, RPG, Shoppers' Stop, Lifestyle, Westside, and Ebony etc. The following data relates to three parameters of the above mentioned retail pack namely turn over (Rs Crore), total floor space (Lakhs Sq. ft) and the total no. of outlets over a period of two years from 2003-2005.

Table 3.34

Retailer	Turnover (Rs Cr)		Total floor space (Lakhs Sq.ft)		Total No. of outlets	
	2003-2004	2004-2005	2003-2004	2004-2005	2003-2004	2004-2005
Pantaloon	650	1300	11.0	30.0	31	74
RPG	545	800	5.2	7.5	110	134
Shoppers' stop	404	545	6.3	8.5	13	40
Lifestyle	230	310	3.2	3.2	7	0
Westside	120	NA	2.3	3.3	14	19
Ebony	85	100	1.7	2.4	8	12

Question:

- (1) Represent the above information with the help of some effective tools to enable a ready and quick assimilation of the facts.
- (2) Using these tools analyze the performance of the retailers in the past two years.

3.6 EXCEL GUIDE **Construction of Frequency Distribution**

Step 1 - Enter the Data in an Excel Spreadsheet

	A
1	40
2	44
3	50
4	52
5	70
6	70
7	72
8	76
9	80
10	80
11	82
12	82
13	82
14	84
15	86
16	88
17	88
18	94
19	94

Step 2 - Create a Table of Classes and BIN Numbers. Bin numbers are the upper boundaries

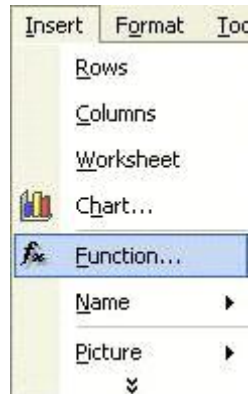
	A	B	C	D	E
1	40				
2	44		Classes	BIN	Frequencies
3	50		40-49	49	
4	52		50-59	59	
5	70		60-69	69	
6	70		70-79	79	
7	72		80-89	89	
8	76		90-100	100	
9	80				
10	80				
11	82				
12	82				
13	82				
14	84				
15	86				
16	88				
17	88				
18	94				
19	94				

Step 3 - Fill in the Frequencies Column Using the Excel Frequency Function

A: Highlight the Frequencies Column

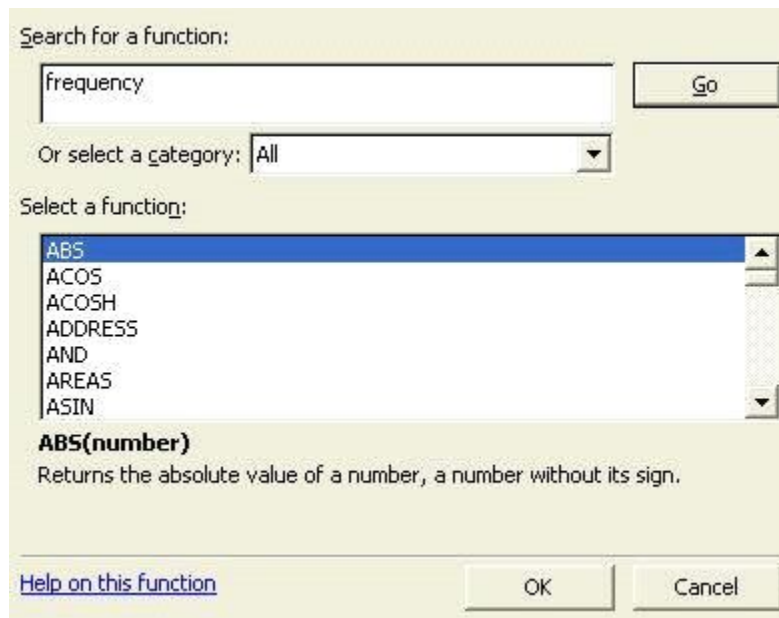
	A	B	C	D	E
1	40				
2	44		Classes	BIN	Frequencies
3	50		40-49	49	
4	52		50-59	59	
5	70		60-69	69	
6	70		70-79	79	
7	72		80-89	89	
8	76		90-100	100	
9	80				
10	80				
11	82				
12	82				
13	82				
14	84				
15	86				
16	88				
17	88				
18	94				
19	94				

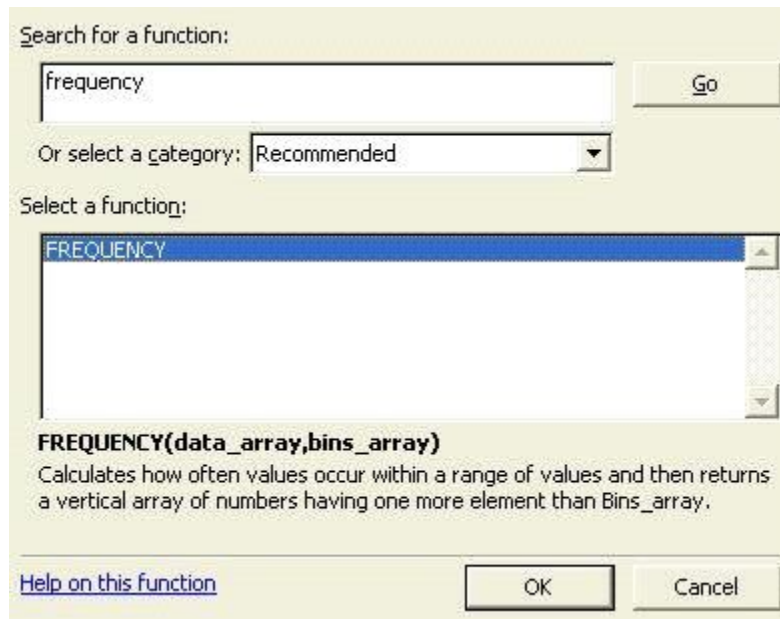
B: Select Insert Function from the main Toolbar



C. Select a category All. In the right side select the function

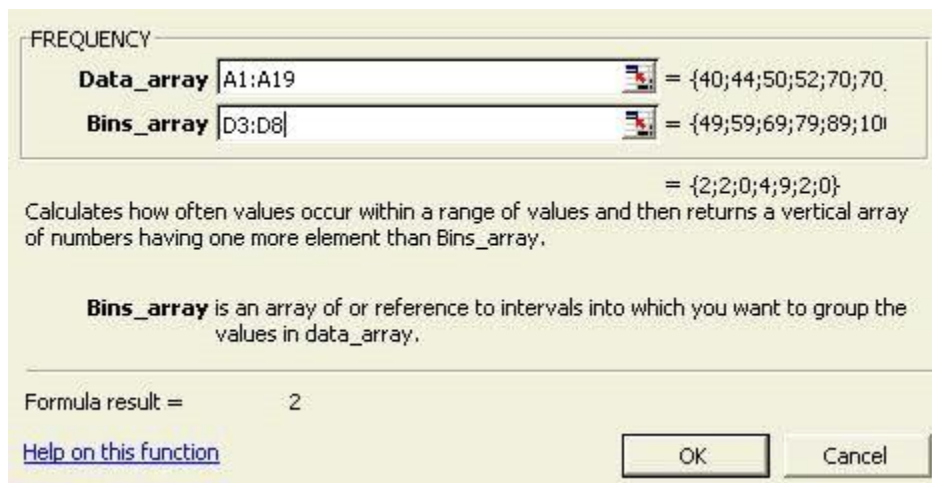
Frequency





D: Click on OK

E: Fill in the Data_array and Bins_array Values



F. To generate the frequencies press Ctrl+Shift+Enter

	A	B	C	D	E
1	40				
2	44		Classes	BIN	Frequencies
3	50		40-49	49	2
4	52		50-59	59	2
5	70		60-69	69	0
6	70		70-79	79	4
7	72		80-89	89	9
8	76		90-100	100	2
9	80				
10	80				
11	82				
12	82				
13	82				
14	84				
15	86				
16	88				
17	88				
18	94				
19	94				

3.7 EXERCISES

- 3.1 Define primary and secondary data. Describe the various methods of collecting primary data.
- 3.2 What do you mean by a questionnaire? State the essential points to be observed in drafting a good questionnaire.
- 3.3 Explain the various diagrams and graphs that can be used for charting a frequency distribution
- 3.4 What is a frequency distribution? What considerations should be given in selecting the class intervals while preparing a frequency distribution?
- 3.5 What is cumulative frequency graph? Explain its uses. Discuss the method of constructing cumulative frequency graph.
- 3.6 K.K. Dey Chem Ltd is a leading Chemical industry. The exports and imports of the company are given below. Draw a suitable diagram to represent the data.

Year	Export (Rs Lakhs)	Import (Rs Lakhs)
1996-97	6500	9659
1997-98	6785	1200
1998-99	7001	1542
1999-2000	8452	1345
2000-2001	8996	1444
2001-2002	9623	1574
2002-2003	9714	1854
2003-2004	1088	1899

3.7 Consider the following data related to cost and profit (Rs lakhs) of two firms I and II. Represent the data with a suitable bar graph.

Item	Firm I	Firm II
Raw material	50	80
Labour	20	90
Overhead	20	50
Profits	10	40

3.8 Explain the method of constructing a pie chart and draw it for the following data.

Export of Cotton	USA	India	Egypt	Brazil	Argentina
(1000 bales)	6367	2999	1688	650	202

3.9 Illustrate the following data of expenditure of an middle class family by a suitable diagram.

Item of expenditure	Percentage of total expenditure
Food	50
Clothing	12
Housing	17
Fuel	8
Education	10
Miscellaneous	3

3.10 Arrange the following data into an ordered array

111	160	159	254	116	841	296	135
120	124	148	225	132	174	212	145
152	154	741	223	134	114	252	125
123	254	120	241	152	452	325	325
154	325	120	142	620	251	201	263

3.11 The profits (Rs lakhs) of 45 companies are given below

Profit	12	27	15	28	40	42	35	37	43
(Rs lakhs)	65	62	53	29	64	69	36	25	18
of 45	55	35	43	26	21	48	43	50	67
companies	23	59	34	68	22	41	42	43	52
	26	37	26	49	53	40	20	18	17

Form a frequency distribution selecting suitable class intervals.

3.12 The marks of statistics obtained by 50 student of an institute is given as follows:

Marks	84	51	65	60	82	64	53	65	55	63
obtained	65	87	67	64	50	69	74	55	65	68
By 50	43	54	77	67	97	66	81	78	78	62
students	37	87	75	98	83	46	59	41	41	74
	90	70	82	79	67	64	50	55	55	54

Using this data construct a frequency distribution table.

3.13 The manager of a company wants to know the ages of the 50 employees engaged in a particular department of his company. The collected data in the raw form is presented below.

Ages of 50	25	32	28	25	36	20	22	41	26	27
employees of	27	35	26	27	37	24	25	43	23	41
manufacturing	28	39	33	33	27	26	28	49	34	36
department	29	40	32	39	29	29	33	46	38	28
	30	21	31	41	30	37	36	25	39	25

Classify the above data selecting suitable class interval.

3.14 What is a histogram? Draw a histogram from the following data:

Class limits	Frequency
5-10	4
10-20	7
20-30	8
30-40	3

3.15 The data related to sales of 100 companies is given below:

Sales (Rs lakhs)	No. of Companies
5-10	5
10-15	12
15-20	13
20-25	20
25-30	18
30-35	15
35-40	10
40-45	7

- (i) Draw less than and more than ogive.
- (ii) Determine the number of companies whose sales are (a) less than Rs. 15 lakhs (b) more than 40 lakhs and (c) between Rs. 15 lakhs and Rs. 36 lakhs.



4

Measures of Central Tendency and Variation



Structure

- 4.1 Introduction
 - 4.1.1 Concept of Central Tendency
- 4.2 Measures of Central Tendency
 - 4.2.1 Arithmetic Mean
 - 4.2.1.1 Properties of Arithmetic Mean
 - 4.2.1.2 Calculation of A.M from Ungrouped Frequency Distribution
 - 4.2.1.3 Calculation of A.M from Grouped Frequency Distribution
 - 4.2.1.4 Calculation of Weighted Mean
 - 4.2.1.5 Correction of Incorrect Observation
 - 4.2.1.6 Mean of Composite Group
 - 4.2.2 Harmonic Mean
 - 4.2.3 Geometric Mean
 - 4.2.3.1 Calculation of Geometric Mean Using Logarithms
 - 4.2.3.2 Combined Geometric Mean
 - 4.2.3.3 Weighted Geometric Mean
 - 4.2.4 Median
 - 4.2.4.1 Computation of Median
 - 4.2.4.2 Quartiles, Deciles and Percentiles
 - 4.2.4.3 Locating Quartile, Deciles and Percentiles Graphically
 - 4.2.5 Mode
 - 4.2.5.1 Computation of Mode
 - 4.2.6 Comparing the Mean, the Median and the Mode
- 4.3 Concept of Variation
- 4.4 Absolute Measures of Variation
 - 4.4.1 Range
 - 4.4.2 Quartile Deviation
 - 4.4.3 Mean Deviation (MD) on Mean Absolute Deviation (MAD)
 - 4.4.4 Standard Deviation
 - 4.4.4.1 Important Properties of SD
 - 4.4.4.2 Calculation of Standard Deviation
 - 4.4.4.3 Combined Standard Deviation
- 4.5 Relative Measures of Variation
 - 4.5.1 Coefficient of Variation
 - 4.5.2 Coefficient of Quartile Deviation
 - 4.5.3 Coefficient of Mean Deviation
- 4.6 Skewness
- 4.7 Kurtosis
- 4.8 Caselets
- 4.9 Excel Guide
- 4.10 Exercises

4.1 INTRODUCTION

In Chapter 3 our aim was to describe how raw data could be collected and arranged in a proper and precise form to make it easily understandable and manageable. Now in the subsequent chapters including the present, the focus will be on the exploration of the properties of numerical data. So in the present chapter, first the students will be introduced to the concept and different measures of central tendency. Secondly, they will also learn the measures of variation.

4.1.1 Concept of Central Tendency

One of the most important objectives of statistical analysis is to determine a central most or average value of the observation that will represent the whole set of observation. Central Tendency may be defined as the parameter in a series of statistical observation, which reflects a central value of the same series. In other word the major characteristics of an entire series of data reflected by a parameter called central tendency.

A few important definitions of central tendency are:

“A measure of central tendency is a typical value around which other figures congregate”

Simpson & Kafka

“An average is a single value within the range of the data that is used to represent all the values in the series. Since an average is somewhere within the range of data it is sometimes called a measure of central value.”

Croxton and Cowden

Desirable Properties of a Measure of Central Tendency

A good measure of central tendency must possess the following characteristics:

- (1) It should be well defined.
- (2) It should be easy to compute.
- (3) It should be easy to understand.
- (4) It should be based on all observations.
- (5) It should be capable of further mathematical treatment.
- (6) It should not be affected by extreme observations.
- (7) It should not be affected much by fluctuations of sampling.

4.2 MEASURES OF CENTRAL TENDENCY

Basically there are three measures of central tendency – Mean, Median and Mode. Mean again is of three types- Arithmetic Mean (AM), Geometric Mean (GM) and Harmonic Mean (HM).

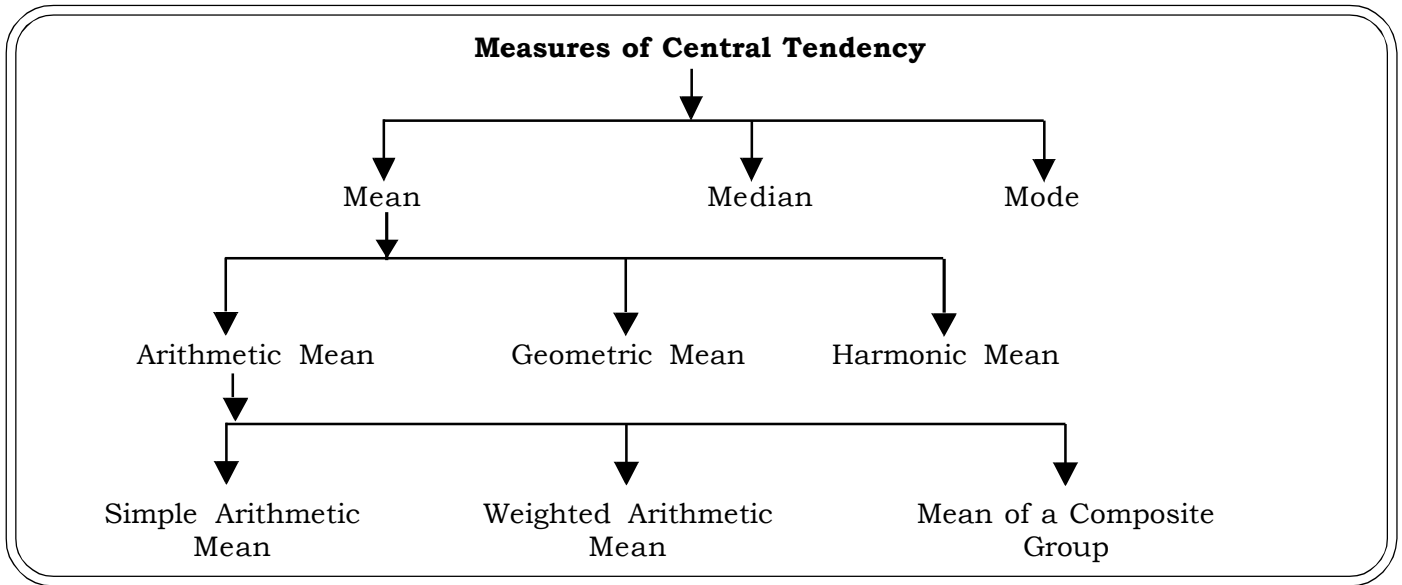


Figure 4.1

Measures of Central Tendency

4.2.1 Arithmetic Mean (AM):

Arithmetic mean (AM) of a set of observations is defined as the sum of the observations divided by their number. A.M calculated from a distribution without frequency is termed as *Simple A.M* and is defined as follows:

If X_1, X_2, \dots, X_n observation are there, then in mathematical form,

$$\text{Simple AM or } \bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

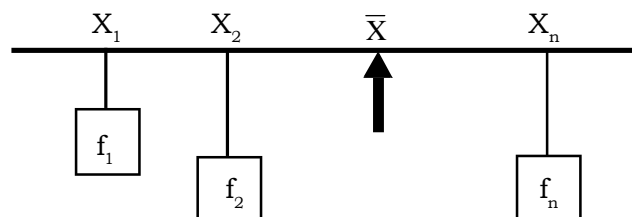
$$= \frac{1}{n} \sum_{i=1}^n X_i \quad i = 1, 2, \dots, n$$

where n is the number of observations.

For example A.M of 3, 4, 5, 7,8,10,12 is $49/7 = 7$

And if the corresponding frequencies of X_1, X_2, \dots, X_n are given as f_1, f_2, \dots, f_n respectively, then the arithmetic mean is defined as

$$\bar{X} = \frac{1}{N} \sum_{i=1}^n f_i X_i \quad \text{where } N = \sum_{i=1}^n f_i = \text{Total frequency}$$



Weighted arithmetic mean takes into account the importance of each value to the overall data with the help of the weights. Frequency i.e. the number of occurrence indicates the relative importance of a particular data in a group of observation.

For example consider n observations x_1, x_2, \dots, x_n each with weights w_1, w_2, \dots, w_n respectively. Then the weighted AM of the n observations are

$$\bar{x}_w = \frac{x_1 w_1 + x_2 w_2 + \dots + x_n w_n}{w_1 + w_2 + \dots + w_n}$$

A weighted mean may be used in case the relative importance of each observation differs or when rates, percentages or ratios are being averaged.

Although the Arithmetic Mean is not the only 'Mean' (there are also Geometric Mean, Harmonic Mean), it is by far the most commonly used. Therefore, if the term 'Mean' is used without specifying whether it is the Arithmetic Mean, the Geometric mean, or Harmonic Mean, it is assumed to refer to the Arithmetic Mean. In general, Arithmetic Mean of X_1, X_2, \dots, X_n is denoted by \bar{X} .

A.M is considered to be the **best measure of Central Tendency** as its computation is based on each and every observation. It is quite easy to calculate and also easy to understand. But the great disadvantage of this measure is that it is highly affected by the extreme values. For Example A.M of 3, 4, 5, 7, 8, 10, 12 is 7, but A.M of 3, 4, 5, 7, 8, 10, 12, 63 is 14, which is just double of the previous A.M. Also, AM cannot be calculated for a grouped frequency distribution with open-end classes.

4.2.1.1 Properties of Arithmetic Mean:

- (i) The sum of a set of observation is equal to the product of their number and Arithmetic

$$\text{Mean i.e. } \sum_{i=1}^n X_i = n\bar{X}$$

Proof:

$$\bar{X} = \frac{\sum X_i}{n}$$

or $\boxed{\sum_{i=1}^n X_i = n\bar{X}}$

- (ii) Sum of the deviations of a set of observations says X_1, X_2, \dots, X_n from their mean \bar{X} is equal to zero. That is:

$$\sum_{i=1}^n (X_i - \bar{X}) = 0$$

Proof:

$$\begin{aligned} & \sum_{i=1}^n (X_i - \bar{X}) \\ &= \sum_{i=1}^n X_i - \sum_{i=1}^n \bar{X} \end{aligned}$$

$$= n\bar{X} - n\bar{X}$$

$$= 0$$

Thus $\boxed{\sum_{i=1}^n (X_i - \bar{X}) = 0}$

(iii) The sum of the squares of the deviations of a set of observation from any number, say A, is the least when A is \bar{X} .

Proof: $\sum_{i=1}^n (X_i - A)^2$ will be minimum when its first order derivative with respect to A will be zero.

$$\text{Thus, } -2 \sum_{i=1}^n (X_i - A) = 0$$

$$\text{or } \sum_{i=1}^n (X_i - A) = 0$$

$$\text{or } \sum_{i=1}^n X_i - \sum_{i=1}^n A = 0$$

$$\text{or } \sum_{i=1}^n X_i = \sum_{i=1}^n A$$

$$\text{or } \sum_{i=1}^n X_i = nA$$

$$\text{or } \frac{\sum X_i}{n} = \frac{nA}{n}$$

$$\text{or } \bar{X} = A$$

$\boxed{\text{Thus } \sum (X_i - A)^2 \text{ is the least when } A = \bar{X}}$

(iv) Arithmetic mean is dependent on both change in origin and change in scale.

Proof:

Let X_i be the original variable. After changing its origin and scale respectively by A and d, let the new variable obtained be Y_i and it can be defined as

$$Y_i = \frac{X_i - A}{d}$$

$$\text{or } X_i = A + dY_i$$

$$\text{or } \sum X_i = nA + d \sum Y_i$$

$$\text{or } \frac{\sum X_i}{n} = \frac{nA}{n} + d \frac{\sum Y_i}{n}$$

$$\text{or } \bar{X} = A + d\bar{Y}$$

Arithmetic mean is dependent on both change in origin and change in scale

4.2.1.2 Calculation of A.M. from Ungrouped Frequency Distribution

Let X_1, X_2, \dots, X_n be n observations with respective frequencies f_1, f_2, \dots, f_n respectively.

Then the arithmetic mean of these observations is given by:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^n X_i f_i$$

$$\text{where } N = \sum_{i=1}^n f_i$$

= Total Number of observations

Example 4.1: The production manager of a company making shampoo has purchased a machine to fill 500 litre shampoo bottles within 499 & 502 litres. To test the machine, a sample of 10 units are taken and the measurements are as follows:

498.1, 499.2, 501.4, 502, 500, 499, 500.2, 499.5, 500.7, 501.2

Is it necessary to recalibrate the machine or is it well within the specifications?

Solution:

$$\text{The mean of the measurements} = \frac{5001.3}{10} = 500.13$$

The mean is well within specifications. There is no need to recalibrate the machine, as far as the mean of the process is concerned. However, further studies regarding variability of the process need to be conducted.

Example 4.2: Calculate arithmetic mean from the following frequency distribution of marks at a test in statistics

Frequency Distribution of Marks of Students

Marks	No. of student
25	2
30	3
35	4
40	8
45	9
50	4
55	3
60	2

Solution:

Calculation of mean marks

X_i	f_i	$X_i f_i$
25	2	50
30	3	90
35	4	140
40	8	320
45	9	405
50	4	200
55	3	165
60	2	120
Total	35	1490

In our sample, total number of observations (N) = 35, $\sum X_i f_i = 1490$

$$\therefore \text{The arithmetic mean} = \frac{\sum X_i f_i}{N} = \frac{1490}{35} = 42.57.$$

Thus, mean marks of the students = 42.57.

Example 4.3: The number of patients admitted in a hospital monitored over a period of 20 days is

45	80	85	70	75
60	78	79	83	90
85	82	71	70	77
82	85	87	92	65

Find the average number of patients admitted in the hospital.

Solution:

The average number of patients admitted in the hospital:

$$\begin{aligned} & \frac{\sum \mathbf{x}}{\mathbf{n}} \\ &= \frac{1541}{20} = 77.05 \\ &\cong 77 \end{aligned}$$

Example 4.4: The number of customers arriving at a railway reservation counter on 7 days of a week are 160, 140, 130, 90, 100, 80, 95 (in “000”s).

Find the weekly average number of customers.

Solution:

$$\text{The weekly average} = \frac{795000}{7} = 113571 \text{ customers}$$

4.2.1.3 Calculation of A.M from Grouped Frequency Distribution

As we know, a grouped frequency distribution consists of data that are grouped by classes. For calculating A.M from this type of distribution, at the first step the mid values are calculated. Then they are multiplied by the respective class frequencies and the results are added up. Finally dividing the sum by the total number of observations the A.M is calculated.

Let x_1, x_2, \dots, x_n be the mid values of n class intervals.

Let f_1, f_2, \dots, f_n be the respective frequencies.

Then, the arithmetic mean is:

$$\begin{aligned} \bar{\mathbf{X}} &= \frac{1}{\sum_{i=1}^n \mathbf{f}_i} \sum_{i=1}^n \mathbf{x}_i \mathbf{f}_i \\ &= \frac{1}{\mathbf{N}} \sum_{i=1}^n x_i f_i, \quad \text{where } \mathbf{N} = \sum_{i=1}^n \mathbf{f}_i \end{aligned}$$

Example 4.5: Calculate the mean annual tax payment by 56 managers from the following distribution:

Annual Tax Payment

Annual Tax Paid (Rs Thousand)	No. of Managers
5 - 10	5
10 - 15	8
15 - 20	10
20 - 25	15
25 - 30	9
30 - 35	6
35 - 40	3

Solution:

Calculation of Annual Average Tax

Annual Tax Paid (Rs Thousand)	No. of Manager (f_i)	Mid Values (X_i)	$f_i X_i$
5 - 10	5	$(5 + 10)/2 = 7.5$	37.5
10 - 15	8	$(10 + 15)/2 = 12.5$	100
15 - 20	10	$(15 + 20)/2 = 17.5$	175
20 - 25	15	$(20 + 25)/2 = 22.5$	337.5
25 - 30	9	$(25 + 30)/2 = 27.5$	247.5
30 - 35	6	$(30 + 35)/2 = 32.5$	195
35 - 40	3	$(35 + 40)/2 = 37.5$	112.5
	$N = \sum f_i = 56$		$\sum f_i X_i = 1205$

$$\text{Mean } \bar{X} = \frac{\sum f_i X_i}{N} = \frac{1205}{56} = 21.52$$

The annual average tax paid by a manager is Rs. 21.52 thousand.

Example 4.6: The following frequency distribution represents the time taken in seconds to serve customers at a fast food take away. Calculate the mean time taken by to serve customers.

Time Taken to Serve Customers at a fast food takeaway

Time Taken (in seconds)	Frequencies
40 - 60	6
60 - 80	12
80 - 100	15
100 - 120	12
120 - 140	10
140 - 150	5

Solution:

Calculation of Mean time taken to serve customers

Classes	f	x	xf
40 - 60	6	50	300
60 - 80	12	70	840
80 - 100	15	90	1350
100 - 120	12	110	1320
120 - 140	10	130	1300
140 - 150	5	150	750

Thus, mean time taken to serve customers is

$$= \frac{\sum xf}{\sum f} = \frac{5860}{60} = 97.6 \text{ seconds}$$

Short cut method of calculating A.M

By applying the properties of scale and origin change, the A.M can be calculated more easily. Example 4.7 is an illustration of how to calculate the A.M from a grouped frequency distribution, by using a simplified method.

Example 4.7: Find the arithmetic mean of the frequency distribution of monthly wages of 100 labourers of a coalmine.

Monthly wages of Labourers of a coalmine

Wages (00Rs)	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of labourers	6	16	27	23	16	10	2

Solution:

Let $A = 45$ and $d = 10$, then we make the transformation: $Y_i = \frac{X_i - 45}{10}$.

Calculation of mean monthly wages of 100 labourers

Class interval	Midpoints (X_i)	Frequency (f_i)	$Y_i = \frac{X_i - A}{d}$	$Y_i f_i$
10 - 20	15	6	-3	-18
20 - 30	25	16	-2	-32
30 - 40	35	27	-1	-27
40 - 50	45	23	0	0
50 - 60	55	16	1	16
60 - 70	65	10	2	20
70 - 80	75	2	3	6
		100		-35

Here total number of observation is 100. $A = 45$, $d = 10$

\therefore The arithmetic mean of the original observations is:

$$\begin{aligned}\bar{X} &= A + d\bar{Y} \\ &= 45 + 10 \frac{-35}{100} \\ &= 45 - 3.5 \\ &= 41.5\end{aligned}$$

Example 4.8: A company manufacturers polythene bags. The bags are evaluated on the basis of their strength by buyers. The strength depends on their bursting pressures. The following data relates to the bursting pressures recorded in a sample of 90 bags. Find the average bursting pressure.

Bursting Pressure in a sample of 90 bags

Bursting Pressure	No. of Bags
5 - 10	10
10 - 15	15
15 - 20	20
20 - 25	25
25 - 30	20
Total	90

Solution:**Calculation of average bursting pressure**

Bursting Pressure	x	f	$u = \frac{x-20}{5}$	fu
5 - 10	7.5	10	-2	-20
10 - 15	12.5	15	-1	-15
15 - 20	17.5	20	0	0
20 - 25	22.5	25	1	25
25 - 30	27.5	20	2	40
			30	

$$\bar{x} = 20 + 5 \times \frac{30}{90} = 20 + 1.67 = 21.67$$

4.2.1.4 Calculation of Weighted Mean

Example 4.9: 5 students of a B.Sc. (Hons) course are marked by using the following weighing scheme:

Mid-Term - 20%

Project - 10 %

Attendance - 10%

Final Exam - 60%

The marks of the students in the various components are

Marks of students

Student No.	Mid-Term	Project	Attendance	Final Exam
1	65	70	80	80
2	48	58	54	60
3	58	63	65	50
4	58	70	54	60
5	60	65	70	70

Calculate the average marks in the examination

Solution:

The calculation of the final marks would be as follows:

For student No. 1

$$\begin{aligned} \text{Final Mark} &= \frac{20 \times 65 + 10 \times 70 + 10 \times 80 + 60 \times 80}{20 + 10 + 10 + 60} \\ &= 76 \end{aligned}$$

The Rest of the calculations are in the following table

Average Marks of Students

Student No.	Mid-Term	Project	Attendance	Final Marks	Average Marks
1	65	70	80	80	76
2	48	58	54	60	56.8
3	58	63	65	50	54.4
4	58	70	54	60	60
5	60	65	70	70	67.5

4.2.1.5 Correction of Incorrect Observation

Example 4.10: The arithmetic mean of 50 students was given 44. But later on it was found that marks of a student which was read wrongly as 54 was actually 34. Now, correctly calculate the mean.

Solution:

$$\sum X = N\bar{X} = 50 \times 44 = 2200$$

The incorrect figure is 54 and the correct figure is 34.

Subtract the incorrect figure from $\sum X$ and add the correct figure to the same.

$$\text{Thus corrected } \sum X = 2200 - 54 + 34 = 2180$$

$$\text{Hence correct average} = \frac{2180}{50} = 43.6$$

Example 4.11: The numbers 3.2, 5.8, 7.9 and 4.5 have frequencies X , $X + 2$, $X - 3$ and $X + 6$ respectively. If their arithmetic mean is 4.876, find the value of X .

Solution:**Calculation of frequencies given the mean**

Numbers (X_i)	Frequency (f_i)	$f_i X_i$
3.2	X	$3.2X$
5.8	$X + 2$	$5.8(X + 2)$
7.9	$X - 3$	$7.9(X - 3)$
4.5	$X + 6$	$4.5(X + 6)$
	$N = \sum f_i = 4X + 5$	$\sum f_i X_i = 21.4X + 14.9$

$$\text{The mean} = \bar{X} = \frac{\sum X_i f_i}{N}$$

$$\text{or, } 4.876 = \frac{21.4X + 14.9}{4X + 5}$$

$$\text{On simplifying, } X = 5$$

4.2.1.6 Mean of Composite Group

If two groups contain respectively n_1 and n_2 observations with mean \bar{X}_1 and \bar{X}_2 respectively then the combined mean (\bar{X}) of the combined group of $N_1 + N_2$ observations is given by:

$$\bar{X} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2}$$

This formula can be extended for any number of observations.

In general for k groups, the combined/composite mean of the k groups is

$$\bar{X} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2 + \dots + n_k \bar{X}_k}{n_1 + n_2 + \dots + n_k}$$

Where

\bar{X}_1 → Mean of the first group

.

.

.

\bar{X}_k → Mean of the k^{th} group

n_1 → Size of the first group

n_2 → Size of the second group

n_k → Size of the k^{th} group

Example 4.12: The mean weight of 150 students in a class is 60 kg. The mean weight of boys in the class is 70 kg and that of girls is 55 kg. Find the number of boys and number of girls in the class.

Solution:

Let the number of boys = N_1

Number of girls = N_2

$$\text{Thus } N_1 + N_2 = 150 \quad \dots (1)$$

The mean weight of all the students = $\bar{X} = 60$ kgs

The mean weight of the boys = $\bar{X}_1 = 70$ kgs

The mean weight of the girls = $\bar{X}_2 = 55$ kgs

From the formula of composite mean we can write the mean of the two groups as

$$\bar{X} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2}$$

$$\text{or, } 60 = \frac{N_1 70 + N_2 55}{150}$$

$$\text{or, } 70N_1 + 55N_2 = 9000$$

$$\text{or, } 14N_1 + 11N_2 = 1800 \quad \dots (2)$$

Now by solving equation (1) and (2)

$$\cancel{14}N_1 + 14N_2 = 2100$$

$$\cancel{14}N_1 + 11N_2 = 1800$$

$$\hline$$

$$3N_2 = 300$$

$$N_2 = 100$$

Therefore, $N_1 = 150 - 100 = 50$

∴ The number of boys is 50 and number of girls is 100.

Example 4.13: A factory has 3 shifts: morning, evening and night shift. The morning shift has 200 workers, the evening shift has 150 workers and the night shift has 100 workers. The mean wage of the morning shift workers is Rs.200, the evening shift workers is Rs.180 and the overall mean of the workers is Rs.160. Find the mean wage of the night shift workers.

Solution:

Let

\bar{x}_1 = mean wage of the morning shift workers

n_1 = total number of morning shift workers

\bar{x}_2 = mean wage of the evening shift workers

n_2 = total number of evening shift workers

\bar{x}_3 = mean wage of the night shift workers

n_3 = total number of night shift workers

\bar{x} = overall mean wage of all the workers (3 shifts combined)

Given,

$$\bar{x}_1 = \text{Rs.}200 \qquad n_1 = 200$$

$$\bar{x}_2 = \text{Rs.}180 \qquad n_2 = 150$$

$$\bar{x}_3 = \text{To be calculated} \qquad n_3 = 100$$

$$\bar{x} = \text{Rs.}160$$

The Composite mean wage of all the workers is:

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3}$$

$$\Rightarrow 160 = \frac{200 \times 200 + 150 \times 180 + 100\bar{x}_3}{450}$$

$$= \frac{40000 + 27000 + 100\bar{x}_3}{450}$$

$$\Rightarrow 100\bar{x}_3 = 72000 - 67000$$

$$\bar{x}_3 = 50$$

Thus, Mean wage of the night shift workers = Rs.50

Example 4.14: A professor of statistics decided to grade his students on the following basis:

Class performance (C.P.): 20 percent

Quizzes (Q): 10 percent

Mid Term Examination (MT): 10 percent

Assignments (A): 10 percent

End Term Examination (ET): 50 percent

A student has scored the following marks in each of these criteria:

CP : 70

Q : 80

MT: 80

A : 85

ET : 80

Calculate the weighted average of this student.

Solution:

$$\begin{aligned}\text{Composite Average} &= \frac{20 \times 70 + 10 \times 80 + 10 \times 80 + 10 \times 85 + 50 \times 80}{100} \\ &= \frac{7850}{100} = 78.5\%\end{aligned}$$

4.2.2 Harmonic Mean (HM)

Harmonic mean is another important method of calculating an average. Typically, it is appropriate for situations when the average of rates is desired. Harmonic mean of a set of observations is defined as the reciprocal of the arithmetic mean of the reciprocal of the observations.

For a set of n observations X_1, X_2, \dots, X_n , simple Harmonic Mean is defined as:

$$\text{HM} = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \dots + \frac{1}{X_n}} = \frac{n}{\sum \frac{1}{X_i}}$$

and for the same set of observations with frequencies f_1, f_2, \dots, f_n respectively, the Harmonic Mean is calculated as:

$$\text{HM} = \frac{N}{\frac{f_1}{X_1} + \frac{f_2}{X_2} + \dots + \frac{f_n}{X_n}} = \frac{N}{\sum \frac{f_i}{X_i}}$$

The use of HM is limited. This measure gives the largest weight to the smallest item and the smallest weight to the largest item. Because of this, HM is preferably used when there are a few extremely large or small values. HM is the most appropriate average when calculating the average speed of vehicle when the speed is expressed as X kms (Distance) per Y (Hour).

Example 4.15: A man traveled 10 miles four times. The first time 40 m.p.h, second time 30 m.p.h, third time 35 m.p.h and fourth time 25m.p.h. Calculate the average speed.

Solution:

Here HM will be the best average.

$$\begin{aligned}\text{Average speed} &= \frac{4}{\frac{1}{40} + \frac{1}{30} + \frac{1}{35} + \frac{1}{25}} \\ &= 31.52 \text{ m.p.h.}\end{aligned}$$

Example 4.16: A cycle is running at the rate of 15kms/ hour during the first 30kms, 20 kms/hr during the second 30 kms. Find the average speed of the cycle.

Solution:

We will use the harmonic mean to find the average speed of the cycle

$$H = \frac{1}{\frac{1}{2} \left(\frac{1}{5} + \frac{1}{20} \right)} = \frac{1}{0.5(0.07 + 0.05)} = \frac{1}{0.06}$$

$$= 16.67 \text{ km/hr.}$$

Example 4.17: The profit earned by 19 companies is given below:

Profit (Rs. lakh)	: 20-25	25 - 30	30-65	35-40
No. of Companies	: 4	7	4	4

Calculate the Harmonic mean of profit earned.

Solution: The harmonic mean of profit earned is calculated below:

Harmonic mean of profit earned by companies

Profit (Rs. lakhs)	Mid Value (x_i)	No. of Companies (f_i)	$\frac{1}{x_i}$	$f_i \left(\frac{1}{x_i} \right)$
20 - 25	22.5	4	$\frac{1}{22.5} = 0.04$	0.16
25 - 30	27.5	7	$\frac{1}{27.5} = 0.036$	0.252
30 - 35	32.5	4	$\frac{1}{32.5} = 0.0307$	0.1228
35 - 40	37.5	4	$\frac{1}{37.5} = 0.027$	0.108
		19		0.6428

$$\text{The harmonic mean} = \frac{\sum_{i=1}^n f_i}{\sum_{i=1}^n f_i \left(\frac{1}{x_i} \right)}$$

$$= \frac{19}{0.6428}$$

$$= 29.56$$

Thus, the average profit earned by companies in Rs. 29.56 lakhs.

4.2.3 Geometric Mean (GM)

Geometric mean of a set of n observation is the nth root of their product.

For a simple ungrouped series of n observation viz. X_1, X_2, \dots, X_n simple GM is as follows

$$GM = \sqrt[n]{X_1 \times X_2 \times \dots \times X_n}$$

On the other hand, for a frequency distribution $X_i/f_i, i = 1, 2, \dots, n$ the GM is:

$$GM = \sqrt[N]{X_1^{f_1} \times X_2^{f_2} \times \dots \times X_n^{f_n}}$$

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

A GM is often used to calculate the rate of change of population growth.

GM is also useful in averaging ratios, rates and percentages. The following example shows the rationale of using GM instead of AM in calculating the appropriate saving rate. Besides these, the other important use of GM is in constructing index numbers. However, just like HM, if any of the observation is zero, it is not possible to calculate GM for that observation.

Example 4.18: Let initially we deposit Rs 100 and the interest rate for the six coming years as given in Table 4.15

Growth of Rs. 100 in a saving account

Year	Interest Rate (%)	Growth Factor	Saving at the end Year of the year
1	5	1.05	105
2	6	1.06	111.3
3	7	1.07	119.09
4	8	1.08	128.62
5	9	1.09	140.19
6	10	1.10	154.21

$$\text{Growth factor} = 1 + \frac{\text{interest - rate}}{100}$$

$$\text{The AM of growth factor} = \frac{1.05 + 1.06 + 1.07 + 1.08 + 1.09 + 1.10}{6} = 1.075$$

Using this growth factor, the savings at the end of the sixth year is, $100 \times 1.075 \times 1.075 \times 1.075 \times 1.075 \times 1.075 \times 1.075 = 154.33$

This value is slightly higher than the saving figure given in Table 4.15 (154.21)

Now let us see what happens if the averaging is done by taking GM as an average for the growth factor.

$$\begin{aligned}\text{GM of growth factor} &= \sqrt[6]{1.05 \times 1.06 \times 1.07 \times 1.08 \times 1.09 \times 1.10} \\ &= 1.074864\end{aligned}$$

Here the average growth factor is 1.074864, and savings at the end of six years is

$$100 \times 1.074864 \times 1.074864 \times 1.074864 \times 1.074864 \times 1.074864 \times 1.074864 = 154.21$$

This figure is exactly equal to the value calculated in Table 4. 15 (154.21).

Thus it can be said that GM is always a better measure than AM if the data are in ratio form.

Example 4.19: A pharmaceutical company recorded the following percentage increase in net worth over a period of 6 years.

2001	2002	2003	2004	2005	2006
4%	7%	9%	8%	10%	11%

Find the average percentage increase in net worth of the company over the 6 years.

Solution:

Since we need to find the average percentage increase in net worth geometric mean would be the most suitable average

$$\begin{aligned}\text{GM} &= (4 \times 7 \times 9 \times 8 \times 10 \times 11)^{\frac{1}{6}} \\ &= (221760)^{\frac{1}{6}} \\ &= 7.780016\end{aligned}$$

4.2.3.1 Calculation of Geometric Mean using Logarithms

When the number of observations are large, geometric mean can be calculated easily by using logarithms.

For a set of n observations x_1, x_2, \dots, x_n we defined.

$$\text{GM} = (x_1 \dots x_n)^{1/n}$$

Taking log on both sides,

$$\begin{aligned}\log \text{GM} &= \frac{1}{n} \log (x_1 \dots x_n) \\ &= \frac{1}{n} (\log x_1 + \log x_2 + \dots + \log x_n) \\ &= \frac{1}{n} \sum_{i=1}^n \log x_i\end{aligned}$$

and finally,

$$GM = \text{antilog} \left\{ \frac{1}{n} \sum_{i=1}^n \log x_i \right\}$$

For a set of n observations x_1, x_2, \dots, x_n with frequencies f_1, f_2, \dots, f_n respectively, the GM is defined as

$$GM = (x_1^{f_1} x_2^{f_2} \dots x_n^{f_n})^{1/n}$$

Taking log on both sides

$$\log GM = \frac{1}{n} \{f_1 \log x_1 + f_2 \log x_2 + \dots + f_n \log x_n\}$$

and finally,

$$GM = \text{antilog} \left\{ \frac{1}{n} \sum f_i \log x_i \right\}$$

Example 4.20: A machinery is assumed to depreciate 44 percent in value in the first year, 15 percent in the second year and 10 percent 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?
(Vikram University, MBA, 1991)

Solution: The geometric mean is calculated by using log values in the following table:

Average depreciation using GM

Rate of Depreciation (x_i)	No. of years (f_i)	$\log_{10} x_i$	$f_i \log_{10} x_i$
44%	1	1.643453	1.643453
15%	1	1.176091	1.176091
10%	3	1.000	3
			5.819544

By definition,

$$\begin{aligned} GM &= \text{antilog} \left\{ \frac{1}{n} \sum_{i=1}^n f_i \log x_i \right\} \\ &= \text{antilog} \left\{ \frac{1}{5} \times 5.819544 \right\} \\ &= \text{antilog} \{1.163909\} \\ &= 14.587 \\ &\cong 15\% \end{aligned}$$

Thus, the average percentage of depreciation for the 5 year period is = 15%.

4.2.3.2 Combined Geometric Mean

For two series

If G_1 and G_2 are the GM's of two series of observations of sizes n_1 and n_2 respectively, the geometric mean G of the combined series is given by:

$$\log G = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}$$

For more than two series:

The above result may be generalized to find the combined geometric mean of more than two series.

If G_1, G_2, \dots, G_p are the geometric means of P groups each of sizes n_1, n_2, \dots, n_p respectively, then the GM G , of the combined series of size $(n_1 + n_2 + \dots + n_p)$ is given by :

$$G = \text{Antilog} \left\{ \frac{n_1 \log G_1 + \dots + n_p \log G_p}{n_1 + n_2 + \dots + n_p} \right\}$$

Example 4.21: The GM of two series of sizes 10 and 12 are 12.5 and 10 respectively. Find the combined GM of the 22 observations.

Solution: The combined geometric mean G is given by

$$G = \text{Antilog} \left\{ \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2} \right\}$$

$$n_1 = 10, n_2 = 12, G_1 = 12.5, G_2 = 10$$

$$\therefore G = \text{Antilog} \left\{ \frac{10 \times \log 12.5 + 12 \log 10}{22} \right\}$$

$$= \text{Antilog} \left(\frac{22.9691}{22} \right)$$

$$= \text{Antilog} (1.04405)$$

4.2.3.3 Weighted Geometric Mean

When all the observation are not of equal importance, but are given different weights, the weighted GM is used.

Consider the observations x_1, x_2, \dots, x_n with weights $w_1, w_2, w_3, \dots, w_n$ respectively. Then, the weighted GM of the n observations is defined by

$$\text{GM}(w) = \text{Antilog} \left[\frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \log x_i \right]$$

Example 4.22: The weighted geometric mean of 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.

Solution: Given, $x_1 = 20$, $x_2 = 18$, $x_3 = 12$, $x_4 = 4$

$$\text{GM (w)} = \text{Weighted GM} = 11.75$$

$$w_1 = 1, w_2 = 3, w_3 = 4$$

We have to calculate w_4 by formula for weighted GM

$$\log \text{GM (w)} = \frac{1}{w_1 + w_2 + w_3 + w_4} \{w_1 \log x_1 + w_2 \log x_2 + w_3 \log x_3 + w_4 \log x_4\}$$

$$\log 11.75 = \frac{1}{8 + w_4} \{w_1 \log 20 + w_2 \log 18 + w_3 \log 12 + w_4 \log 4\}$$

$$1.070038 = \frac{1}{(8 + w_4)} \{9.383572 + 0.602059 w_4\}$$

$$\Rightarrow w_4 = 1.759$$

Thus $w_4 \cong 2$

Relationship between AM, GM and HM

1. For any number of observations

$$\boxed{\text{AM} \geq \text{GM} \geq \text{HM}}$$

2. For only two observations

$$\boxed{\text{GM} = \sqrt{\text{AM} \times \text{HM}}}$$

3. If all the observations have same value then

$$\boxed{\text{AM} = \text{GM} = \text{HM}}$$

4.2.4 Median

Median is the middle most value of a distribution, which divides the distribution into two equal parts. Thus, exactly half of the observations of a data set lie above the median and half are below it. Median has several advantages over mean. It is less sensitive to the extreme scores than the mean and this makes it a better measure than the mean for highly skewed distributions. The median income is usually more informative than the mean income, for example. Unlike mean, median can be calculated from open-ended classes unless the median falls in an open-ended class. However, the critical statistical procedure of calculating median from grouped frequency distribution makes this measure less popular than mean. Median calculation also has the disadvantage of concentrating only around the median class and not considering all the observations. Before median calculation the arrangement of the data either in ascending or descending order is a must.

4.2.4.1 Computation of Median

For a simple series

When there is odd number of observations, median is simply the middle number after its arrangement either in ascending or descending order. For example, the median of 2, 5, and 7 is 5. Again when there is an even number of observations, the median is the mean of the two middle numbers. Thus, the median of the numbers 2, 4, 8, 12 is $(4 + 8)/2 = 6$.

When individual observations are given, median may be determined by the following steps:

Step 1: Arrange the observations in ascending or descending order of magnitude.

Step 2: Let n = number of observations

(i) When n is odd

$$\text{Median} = \left(\frac{n+1}{2}\right)^{\text{th}} \text{ observation}$$

(ii) When n is even

$$\text{Median} = \text{Mean of the sizes of } \left(\frac{n}{2}\right)^{\text{th}} \text{ and } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ observation.}$$

For ungrouped frequency distribution

In this case, the data are arranged in order of magnitude.

Step 1: Calculate the cumulative frequencies.

Step 2: $N = \sum f_i$ in this case

(i) When N is odd

$$\text{Median} = \text{Size of the } \left(\frac{N+1}{2}\right)^{\text{th}} \text{ observation}$$

(ii) When N is even

$$\text{Median} = \text{Mean of the sizes of the } \left(\frac{N}{2}\right)^{\text{th}} \text{ and } \left(\frac{N}{2} + 1\right)^{\text{th}} \text{ observations}$$

Example 4.23: Calculate median from the following data:

Observations	Frequency
5	20
6	30
9	40
10	45
12	60

Solution:

Calculation of median for ungrouped frequency distribution

Observations	Frequency	Cumulative Frequency
5	20	20
6	30	50
9	40	90
10	45	135
12	60	195

We first calculate $\frac{N}{2}$.

$\frac{N}{2} = 97.5$ The cumulative frequency just greater than 97.5 is 135.

The observation corresponding to the cumulative frequency 135 is 10. So 10 is the median of the given data.

Calculation of Median from grouped frequency distribution

For grouped frequency distribution median can be calculated by using the following formula:

$$M_d = L_1 + \left(\frac{N/2 - C_f}{f_m} \right) \times i$$

where

L_1 = Lower boundary of the median class

N = Total frequency

C_f = Cumulative frequency up to the class that immediately precedes the median class.

f_m = Frequency of the median class

i = Width of the median class.

To compute the median class, we first calculate $\frac{N}{2}$. Then identify the cumulative frequency just greater than $\frac{N}{2}$, and the class corresponding to this cumulative frequency is the median class.

Example 4.24: Calculate Median from the following frequency distribution.

Class	Frequency
50 - 60	6
60 - 70	9
70 - 80	15
80 - 90	25
90 - 100	13
100 - 110	7
110 - 120	5

Solution:

Calculation of Median

Class	Frequency	Cumulative Frequency
50 - 60	6	6
60 - 70	9	15
70 - 80	15	30
80 - 90	25	55 Median class
90 - 100	13	68
100 - 110	7	75
110 - 120	5	80
$N = \sum f_i = 80$		

Since $N/2 = 80/2 = 40$ is contained in the cumulative frequency against the class interval 80-90, it is the median class.

Alternatively, the cumulative frequency just greater than 40 is 55 and the class corresponding to 55 i.e. 80 - 90 is the median class. With $C_f = 30$, $f_m = 25$, $i = 10$ and $L_1 = 80$ we have

$$M_d = L_1 + \left(\frac{N/2 - C_f}{f_m} \right) \times i$$

$$= 80 + \frac{40 - 30}{25} \times 10 = 84$$

Example 4.25: For the following distribution:

- (i) Find the median class.
- (ii) The median value

Class - Interval	Frequencies
20 - 30	3
30 - 40	6
40 - 50	18
50 - 60	10
60 - 70	5

Solution:

Calculation of median

Class - Interval	Frequencies	Cumulative Frequency
20 - 30	3	3
30 - 40	6	9
40 - 50	18	27 (median class)
50 - 60	10	37
60 - 70	5	42

(i) Here $\frac{N}{2} = 21$

And the cumulative frequency just greater than 21 is 27. Therefore, the median class is 40 - 50.

$$\begin{aligned} \text{(ii) Median} &= 40 + \left(\frac{21-9}{18} \right) 10 \\ &= 46.67 \end{aligned}$$

Example 4.26: Given the following frequency distribution, where median is 46, find the missing frequencies.

Frequency Distribution with missing frequencies

Class - Interval	Frequencies
10 - 20	12
20 - 30	30
30 - 40	x
40 - 50	65 median class
50 - 60	y
60 - 70	25
70 - 80	18
Total	229

Solution:

$$\text{Since } N = \sum f_i = 229$$

$$\Rightarrow x + y = 229 - (12 + 30 + 65 + 25 + 18)$$

$$\Rightarrow x + y = 79 \quad \dots \text{ (i)}$$

The median is given to be 46.

Thus, the median class is 40 - 50

Now, by using the formula for the median, $L_1 = 40$, $\frac{N}{2} = 114.5$, $C_f = 42 + x$, $f_m = 65$, $i = 10$

$$46 = 40 + \frac{114.5 - (42 + x)}{65} \times 10$$

$$\Rightarrow x = 34$$

Substituting in (i), $y = 45$

Thus, the missing frequencies are 34 & 45

4.2.4.2 Quartiles, Deciles and Percentiles

Median divides the observations into two equal parts. Quartiles, Deciles and Percentiles are the other three similar types of measures in this respect. Collectively, these measures are often referred to as positional averages. In particular, these measures are useful to measure qualitative characteristics of a data set. Calculations of these measures are also quite similar to that of median.

Quartiles are those values, which divide the total data into four equal parts. Since three points divide the distribution into four equal parts, there are three quartiles, symbolically Q_1 , Q_2 and Q_3 . The first quartile Q_1 is the value such that 25% of the observations are smaller and 75% of the observations are larger than Q_1 . The second quartile is the median. The third quartile Q_3 is the value such that 75% of the observations are smaller and 25% of the observations are larger.



For grouped frequency distribution the following formula is used for the j^{th} quartile, $j = 1, 2, 3$

$$Q_j = L + \left(\frac{jN/4 - C_f}{f} \right) \times i \quad j = 1, 2, 3$$

where

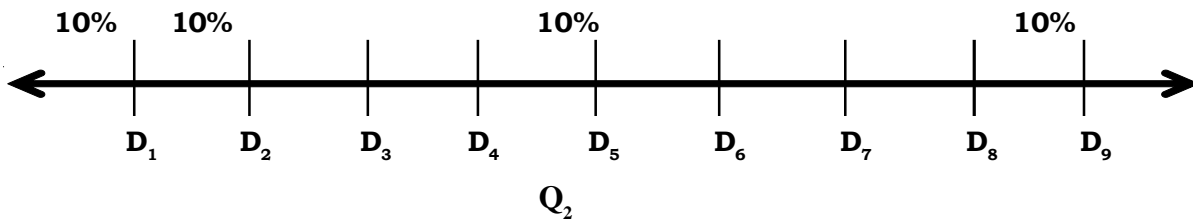
L = lower limit of the quartile class

C_f = cumulative frequency of the class preceding the quartile class

f = frequency of the quartile class,

i is the width of the quartile class.

Deciles are the nine values, which divide the total data into ten equal parts.

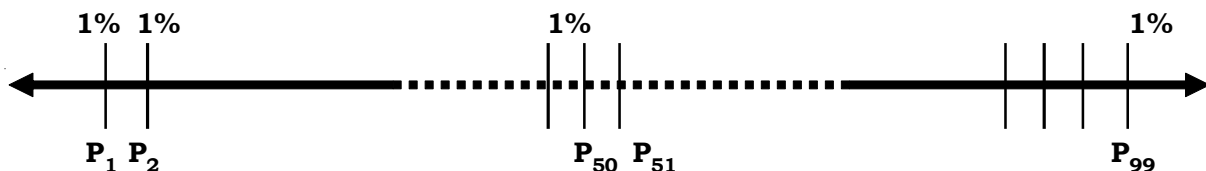


The 9 deciles ($D_1, D_2, D_3, \dots, D_9$) can be calculated by the following formula

$$D_k = L + \left(\frac{kN/10 - C_f}{f} \right) \times i \quad k = 1, 2, \dots, 9. \text{ (k}^{\text{th}} \text{ Decile)}$$

where the symbols have usual meaning.

Percentiles are those values, which divide the total observation into hundred equal parts.



Percentiles are denoted by P_1, P_2, \dots, P_{99} and the i th percentile is calculated by the following formula with notations having their usual meaning.

$$P_l = L + \left(\frac{lN/4 - C_f}{f} \right) \times i \quad l = 1, 2, \dots, 99.$$

To illustrate the computations of quartile, deciles and percentiles consider the following example.

Example 4.27: Calculate Q_1, D_7 and P_{90} from the following grouped data, related to profits of 100 companies in Rs. lakh.

Profit (in Rs. lakh) of Companies

Profit (Rs Laksh)	No. of companies
20 - 30	4
30 - 40	8
40 - 50	18
50 - 60	30
60 - 70	15
70 - 80	10
80 - 90	8
90 - 100	7

Solution:

Calculation of Q_1, D_6 and P_{90}

Profit (Rs Laksh)	No. of companies	Cumulative Frequency
20 - 30	4	4
30 - 40	8	12
40 - 50	18	30 (Quartile class)
50 - 60	30	60
60 - 70	15	75 (Decile class)
70 - 80	10	85
80 - 90	8	93 (Percentile class)
90 - 100	7	100

Calculating first Quartile (Q_1)

Step 1: Locating the quartile class i.e class in which the $N/4$ observation lies

$$\frac{N}{4} \text{th} = \frac{100}{4} \text{th} = 25\text{th Observation.}$$

This lies in the class 40-50. Thus 40-50 becomes the quartile class.

Alternatively, the cumulative frequency just greater than 25 is 30 and the class corresponding to 30 i.e. 40 - 50 is the median class.

Step 2: Calculating Q_1 using the formula

$$\therefore Q_1 = L + \left(\frac{N/4 - C_f}{f} \right) \times i = 40 + \frac{25 - 12}{18} \times 10 = 40 + 7.22 = 47.22$$

Calculating 7th Decile (D_7)

Step 1: Locating the decile class i.e class in which the $iN/10$ observation lies, when $i = 7$

$$\frac{7N}{10} \text{th} = \frac{700}{10} \text{th} = 70\text{th observation.}$$

This lies in the class 60-70 .Thus 60-70 becomes the decile class.

Step 2: Calculating D_7 using the formula

$$\therefore D_7 = L + \left(\frac{7N/10 - C_f}{f} \right) \times i = 60 + \frac{70 - 15}{30} \times 10 = 60 + 18.33 = 78.33$$

Calculating 90th Percentile

Step 1: Locating the percentile class i.e class in which the $90N/100$ observation lies.

$$\frac{90N}{100} \text{th} = \frac{9000}{100} \text{th} = 90\text{th observation which lies in the class 80-90.}$$

Step 2: Calculating P_{90} using the formula

$$\begin{aligned} P_{90} &= L + \left(\frac{\frac{90N}{100} - C_f}{f} \right) \times i \\ &= 80 + \frac{90 - 85}{8} \times 10 \\ &= 80 + 6.25 \\ &= 86.25 \end{aligned}$$

Thus, $Q_1 = \text{Rs. } 47.22$

$D_7 = \text{Rs. } 78.33$

$P_{90} = \text{Rs. } 86.25$

From the above calculations it can be concluded that

- (a) 25% of the companies earn an annual profit of Rs. 47.22 lakh
- (b) 70% of the companies can earn upto Rs. 78.33 lakhs and
- (c) 90% of the companies can earn upto Rs. 86.25 lakh.

Example 4.28: The following data gives marks obtained by 60 students in a marketing class

Class - Interval (CI)	Marks (f)
20 - 40	5
40 - 60	10
60 - 80	30
80 - 100	15
Total	60

- (i) Find Q_1
- (ii) Compute D_2

Solution:

$$(i) Q_1 = L + \left(\frac{\frac{N}{4} - C_f}{f} \right) \times i$$

$$\frac{N}{4} = \frac{60}{4} = 15$$

CI	Marks (f)	x	xf	C
20 - 40	5	30	150	5
40 - 60	10	50	500	15
60 - 80	30	70	2100	45
80 - 100	15	90	1350	60
	60	4100		

The quartile class = 60 - 80

Thus

$$\begin{aligned} Q_1 &= 60 + \frac{15-15}{30} \times 20 \\ &= 60 \end{aligned}$$

Thus, 25% of the students have scores below 60 and 75% of the students have scores above 60

(ii) D_2 : The second decile

$$D_2 = L + \left(\frac{\frac{2N}{4} - C_f}{f} \right) \times i$$

$$\frac{2N}{4} = \frac{2 \times 60}{4} = 30$$

The decile class is 60 – 80

$$\begin{aligned} \text{Thus, } D_2 &= 60 + \frac{30-15}{30} \times 20 \\ &= 60 + 10 = 70 \end{aligned}$$

Thus, 80% of the students have scored above 70

4.2.4.3 Locating Quartile, Deciles and Percentiles Graphically

Step 1: First draw a less than cumulative frequency curve (see Chapter 3), by taking the less than cumulative frequencies on the vertical axis and the upper limit of the class intervals along the horizontal axis.

Step 2: To determine quartiles, deciles and percentiles, plot $i \left(\frac{N}{2} \right)$, $i = 1, 2, 3, \frac{iN}{10}$, $i = 1, 2, \dots, 9$ and $\frac{iN}{100}$, $i = 1, 2, 3, \dots, 90$, as the case may be.

For example to plot Q_1 , we first plot $\frac{N}{2}$ on the X-axis.

Step 3: From this point draw a horizontal line parallel to the X-axis to join the cumulative frequency curve or ogive at a point say Q. From Q, we now draw a perpendicular on the X-axis.

Step 4: The point at which this perpendicular line meets the X-axis is the value of Q_1 in this case (Figure 4.2) or as the case may be.

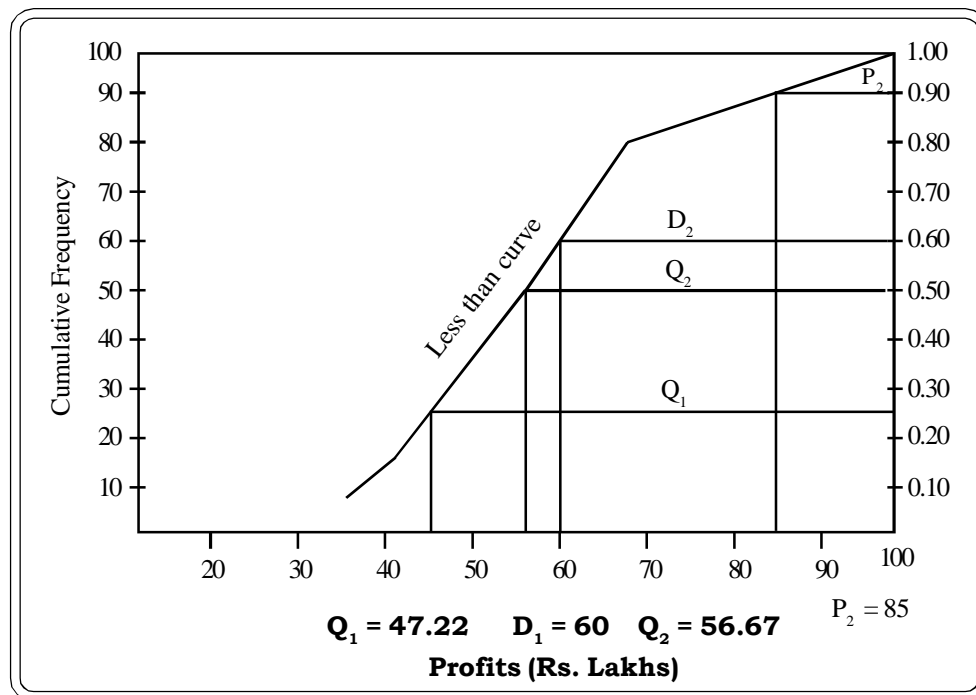


Figure 4.2

Graphical Location of Quartiles, Deciles and Percentiles

Example 4.29: The fuel consumption of 50 household in a locality is given below.

Fuel Consumption: (in litres)	10-20	20-30	30-40	40-50	50-60
Number of Households	5	12	13	12	8

Locate graphically

- (1) The median
- (2) D_5 - The fifth decile

Solution: Step 1: We make the cumulative frequency table as follows:

Cumulative Frequency of Fuel consumption of 50 households

Fuel Consumption C.I.	No. of Households (f)	C.F. (less than)
10 - 20	5	5
20 - 30	12	17
30 - 40	13	30
40 - 50	12	42
50 - 60	8	50

Step 2: We draw the cumulative frequency curve by plotting the C.F.'s on the Y-axis against the upper limits of the class-intervals on the x-axis.

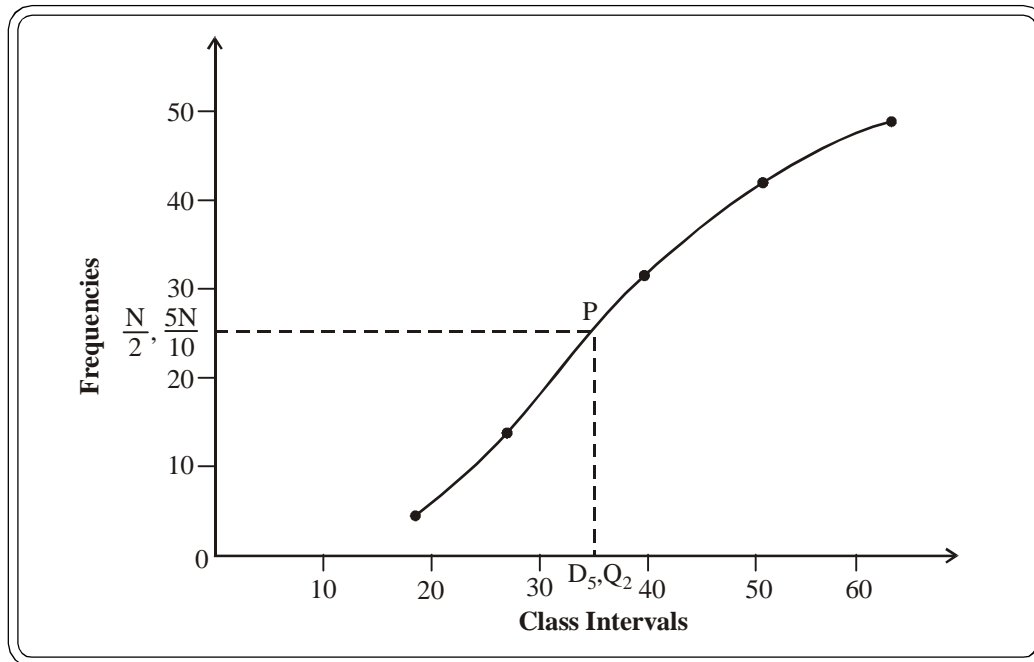


Fig. 4.3

Graphical location of Q_2 and Q_5 .

Step 3: To calculate median, which is also Q_2 , we compute

$$\frac{N}{2} = \frac{50}{2} = 25$$

We plot $\frac{N}{2}$ on the Y-axis

Step 4 : From $\frac{N}{2}$, we draw line parallel to the X-axis to meet the C.F. curve, at the point P.

From P, we draw a line parallel to the Y-axis to meet the X-axis at the point Q_2 , which is the value of the median or the second quartile.

Step 5 : To calculate the 5th Decile we first compute $\frac{5N}{10} = \frac{5 \times 50}{10} = 25$, which is same as $\frac{N}{2}$. The rest of the procedure is same as described in step 4.

From this example we can see that $Q_2 = D_5$.

4.2.5 Mode

The mode is another form of average that can be defined as most frequently occurring value in the data. In a more simplified language mode of a given set of observation is that value which occurs with maximum frequency. Like median, mode is not affected by the extreme values and it

can also be calculated with the open-ended classes. However, like mean and median, this measure is not so popular or reliable as a measure of central tendency because too often in the observation no repetition is observed in the values.

4.2.5.1 Computation of Mode

For a simple series for example 2, 4, 5, 6, 6, 7, 8, 9, 9, 9 the mode is 9 because it occurs maximum number times in the series.

Example 4.30: Series with no mode

Calculate mode for the following annual returns obtained by 10 top-level mutual funds.

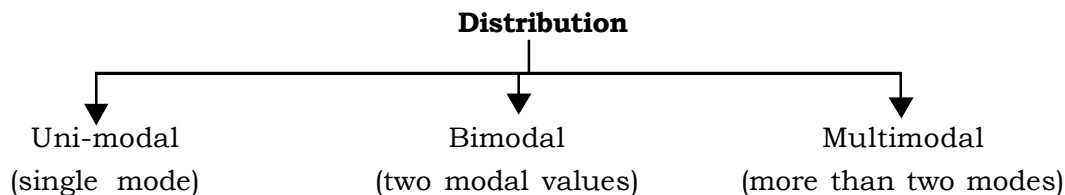
6.2, 6.7, 7.7, 8.5, 9.9, 10.3, 15.4, 16.6, 17.5, 18

Solution:

Here, none of the values are occurring more than once, so there is no mode in the above series.

Example 4.31: Series with more than one mode

A distribution may have more than one mode. A distribution with two modes is called bimodal distribution. For example the series 2, 2, 2, 4, 5, 6, 6, 7, 8, 9, 9, 9 has two modes 2 and 9. And a distribution with more than 2 modes is called a multi-modal distribution. Thus, a modal value is not unique. The following chart shows the different possibilities of modes discussed above.



Calculation of Mode from grouped frequency distribution

In case of grouped frequency distribution, among the classes in the distribution, the class with highest frequency is the modal class. Here, the mode is calculated by the following formula:

$$\text{Mode} = L_{MO} + \frac{d_1}{(d_1 + d_2)} \times i$$

where L_{MO} = Lower boundary of the modal class

d_1 = Difference between frequency of modal class and frequency of the class just preceding the modal class.

d_2 = Difference between frequency of modal class and frequency of the class just following the modal class.

Example 4.32: Calculate mode from the following distribution

Weekly wages (Rs)	Frequency
0 - 100	4
100 - 200	16
200 - 300	60
300 - 400	100
400 - 500	40
500 - 600	6
600 - 700	4

Solution:

For grouped frequency distribution the formula for mode is

$$\text{Mode} = L_{MO} + \frac{d_1}{(d_1 + d_2)} \times i$$

Calculation of mode

Weekly wages (Rs)	Frequency
0 - 100	4
100 - 200	16
200 - 300	60
300 - 400	100 Modal class
400 - 500	40
500 - 600	6
600 - 700	4
Total	230

The class containing the highest frequency i.e. 100 is 300-400. Therefore it is the modal class.

$$L_{MO} = 300, i = 100$$

$$d_1 = (100 - 60) = 40,$$

$$d_2 = (100 - 40) = 60,$$

Thus

$$\begin{aligned} \text{Mode} &= 300 + \frac{40}{40 + 60} \times 100 \\ &= 340 \end{aligned}$$

Example 4.33: The following data is from a shoe store. It give the various shoe sizes sold over a period of one month.

Shoe Size	No. of pairs Sold
10 - 15	100
15 - 20	200
20 - 25	180
25 - 30	150

Find which shoe size sells the most.

Solution:

The mode of this distribution will give the shoe size which has highest sales

$$\text{Mode} = L_{mo} + \frac{d_1}{d_1 + d_2} \times i$$

Calculation of shoe size with the highest sales

Shoe Size	No. of pairs Sold
10 - 15	100
15 - 20	200 (Modal class)
20 - 25	180
25 - 30	150

Modal class (class with the highest frequency) = 15 - 20

$$d_1 = 200 - 100 = 100$$

$$d_2 = 200 - 180 = 20$$

$$\begin{aligned}\text{Mode} &= 15 + \frac{100}{100+20} \times 5 \\ &= 15 + 4.16 \cong 19\end{aligned}$$

Thus, the shoe size which sells the most, is 19.

Example 4.34: A multi national company has set up a new network in its office. To monitor the performance of the network the company recorded the number of server failures over a period of 30 days. The following distribution was obtained:

Number of server failures

No. of Failures (per day)	Frequencies
0 - 2	5
2 - 4	18
4 - 6	7
Total	30

Find the mode.

Solution:

Modal class: 2 - 4

$$d_1 = 18 - 5 = 13$$

$$d_2 = 18 - 7 = 11$$

$$\begin{aligned}\text{Mode} &= 2 + \frac{13}{13+11} \times 2 \\ &= 2 + \frac{26}{24} \\ &= 2 + 1.08 \\ &= 3.08 \\ &\cong 3 \text{ failures per day}\end{aligned}$$

Thus, the maximum number of failures in the office is 3 per day.

4.2.6 Comparing the Mean, the Mode and the Median

The information obtained from these three measures of central tendency in a data distribution is similar in the sense that all reflect some aspect of the data values, which is “typical” of the whole distribution. But they differ in the kind of “typicality” which they report and in how sensitive they are to changes in the values of the observations.

The mean represents the balance point, or center of gravity of the distribution. Its value will change when there is a change in any of the data values in the distribution.

The mode represents the most frequent or probable single value in the distribution. If the value of a data in the distribution changes from a non-modal value to the modal value, the value calculated for the mode remains the same, even though the mean would (and the median might) change. The median represents the middle score of the distribution. If the value of a data is changed so that its position relative to the magnitude of the other values is not changed, the median will remain the same, even though the mean would, and the mode might.

Relationship between Mean, Median and Mode

1. In general, for a slightly asymmetrical distribution:

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

2. For normal or symmetrical distribution:

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

3. For positively skewed distribution:

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

4. For negatively skewed distribution:

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

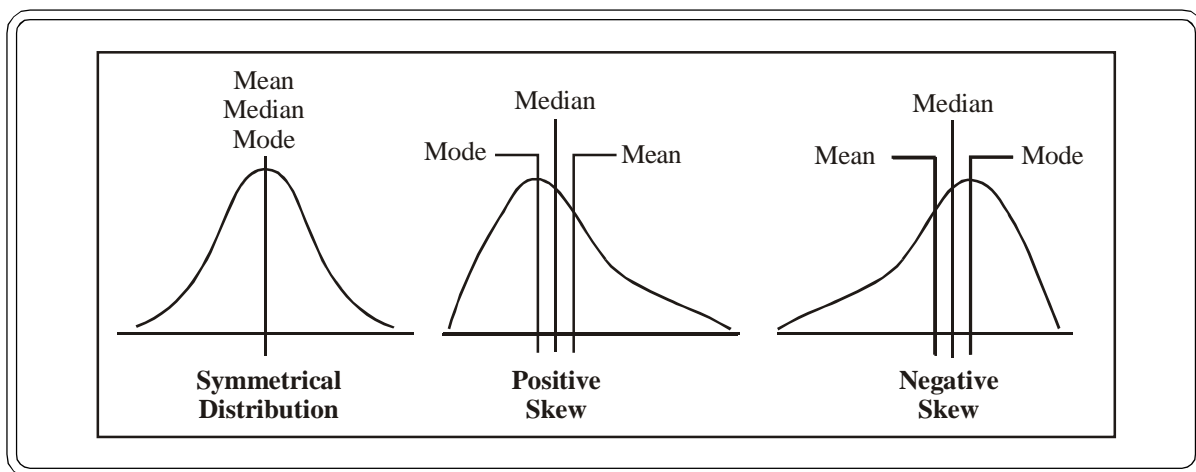


Figure 4.4

Relationship between mean, median and mode

Figure 4.4 shows the relationship between the mean, median and mode for symmetric, positively skewed and negatively skewed distributions respectively.

4.3 CONCEPT OF VARIATION

Measures of central tendency give us good information about the average score in our distribution. However, the information provided by these measures is not sufficient to convey all we need to know about a distribution. For example, we may have very different shapes of distribution with the same central tendency. Consider the following data related to age distribution of two groups A and B:

Table 4.1

Age distribution of two groups

						Average
Group A	22	24	25	26	28	25
Group B	8	15	20	28	54	25

The above-mentioned two groups of observations have the same average i.e 25 years, so we are likely to conclude that the two groups are similar. However, this would be a wrong conclusion as it can be noticed that the observations in group A are close to one another indicating that people in this group are more or less of the age 22years to 28years. While those in Group B are widely dissimilar and have greater variability of ages as it includes a person who is 8 years old on one hand and also a person who is 54 years of age on the other. This clearly indicates that knowledge of the central value alone does not give us a complete picture of the pattern of the distribution. So how the observations are dispersed around the central value is also one of the important matters to judge the quality of data. Measures of dispersion or variability give us the information about the spread of the observations in one distribution. Are the values clustered close together over a small portion of the scale or are the values spread out over a large segment of the scale? In the above example we can say that the dispersion of Group B is more than that of Group A. A proper description of a set of data should include both of these characteristics i.e. the central tendency and variability.

4.4 ABSOLUTE MEASURES OF VARIATION

There are various methods that can be used to measure the variation or dispersion of a data set, each with its own set of advantages and disadvantages. Measures of variation may be either absolute or relative. Measures of absolute variation are expressed in terms of the original units of data. In case of two sets of data, with different units of measurement, the absolute measures are not appropriate. In such cases measures of relative variation or dispersion are used. These measures are independent of units of measurement.

Following are some of the well known measures of variation which provide a numerical index of the variability of the given data:

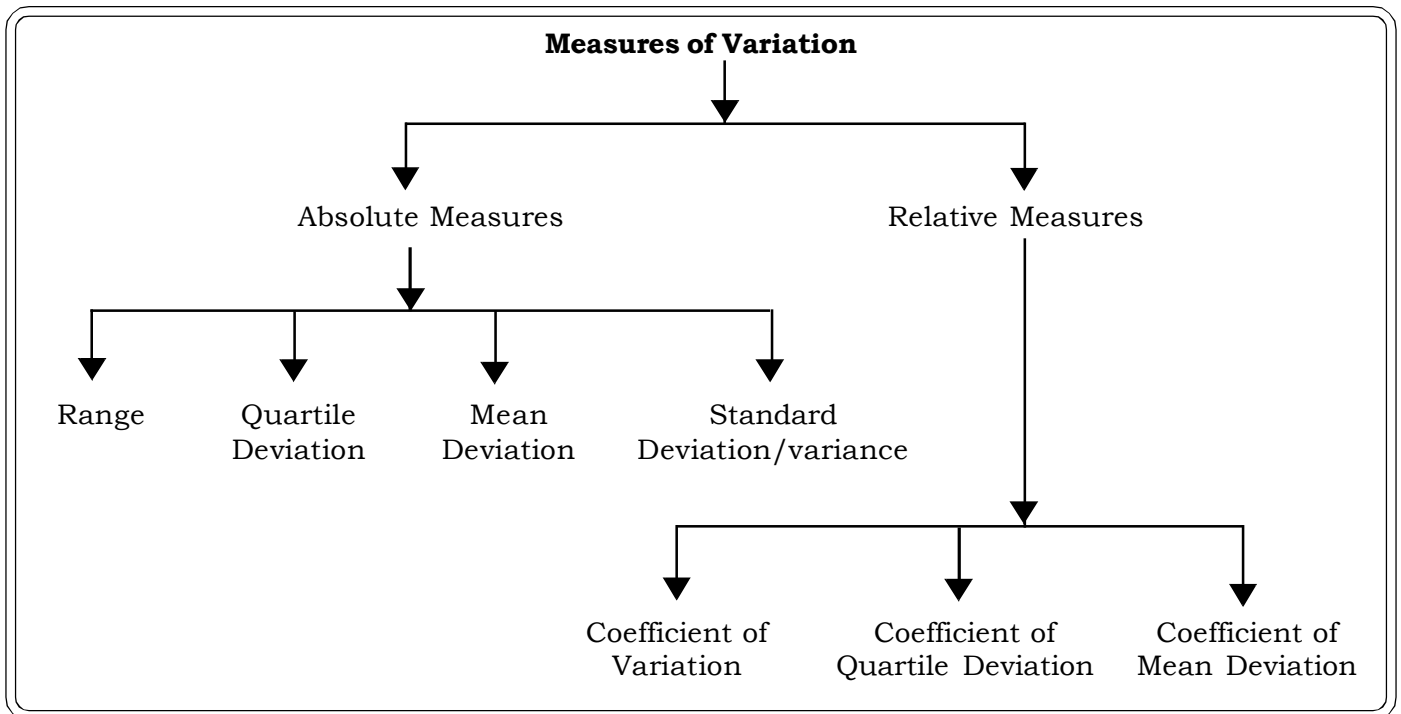


Figure 4.5

Measures of Variation

4.4.1 Range

Range is the preliminary indicator of dispersion. In case of ungrouped data, range is defined as the difference between the highest and lowest values in a distribution. Large range indicates more variability. For example the range of Group A in Table 4.34 is $28 - 22 = 6$ and that of Group B is $54 - 8 = 46$. Thus the variability of Group B is clearly higher than that of Group A as discussed earlier.

In case of grouped data the range is determined by taking the difference between the upper limit of the last class and the lower limit of the first class. For the data of Table 4.29 range is $700 - 0 = 700$. Open-ended frequency distributions have no range.

Range (for un-groped data) = Value of highest observation - Value of lowest observation

Range (for groped data) = Upper limit of the last class - the lower limit of the first class

Range is a simple and easy to understand measure of dispersion. It gives us a rough and ready idea about the variability very quickly. However, because it takes into account only the scores that lie at the two extremes, it is of limited use. This measure considers only the highest and the lowest values of the distribution and fails to take into account the other observations of a data set.

Example 4.35: The following data sets are related to returns achieved by 2 mutual funds. Find the range of each

Mutual fund 1:	10	5	8	4	2
Mutual Fund 2:	12	10	5	9	7

Solution:

Range of mutual fund 1

$$= 10 - 2 = 8$$

Range of mutual fund 2

$$= 15 - 7 = 8$$

The variability in both the sets are same.

4.4.2 Quartile Deviation

Quartile deviation, also known as semi-inter quartile range, is computed by taking the average of the difference between the third quartile and the first quartile. Actually it measures the distance between the lowest and highest of the middle 50 percent of the scores in the distribution. In symbol

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

where Q_1 = first quartile and Q_3 = Second quartile.

The following example illustrates the procedure involved in calculation of the quartiles as well as the quartile deviation.

Example.4.36: Calculate quartile deviation from the following data:

Monthly wages (Rs)	No. of workers
Below 850	12
850 - 900	16
900 - 950	39
950 - 1000	56
1000 - 1050	62
1050 - 1100	75
1100 - 1150	30
1150 and above	10

Solution:

Monthly wages (Rs)	No. of workers.	Cumulative Frequency
Below 850	12	12
850 - 900	16	28
900 - 950	39	67
950 - 1000	56	123
1000 - 1050	62	185
1050 - 1100	75	260
1100 - 1150	30	290
1150 and above	10	300
Total	300	

For Q_1 , $\frac{N}{4}$ th observation = $\frac{300}{4} = 75$ th observation which lies in the class 950-1000.

This is the quartile class.

$$Q_1 = L + \frac{N/4 - C_f}{f} \times h = 950 + \frac{75 - 67}{56} \times 50$$

$$= 950 + 7.14 = 957.14$$

We now have to calculate Q_3 ,

To determine the quartile class, we compute

$$\frac{3N}{4} = \frac{3 \times 300}{4} = 225$$

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

$$\therefore Q_3 = 1050 + \frac{225 - 185}{75} \times 50$$

$$= 1050 + 26.67$$

$$= 1076.67$$

Thus, the quartile deviation is

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

$$= \frac{1076.67 - 957.14}{2}$$

$$= 59.765$$

Quartile deviation is superior to range, as it is not based on two extreme values but rather on middle 50% observation. It can also be calculated from the open-ended classes. However, its unpopularity lies in the fact that just like range, it does not depend on all observations of a data set. The first and last 25% observations remain ignored in this measure.

Example 4.37: A survey was carried out to study how much college students spend on eating out at fast food joints. The following data was collected after questioning 100 college students.

Amount spent by college students on fast food

Expenditure (in Rs.)	No. of students
200 - 400	5
400 - 600	19
600 - 800	30
800 - 1000	26
1000 - 1200	20

Compute the first quartile and give the conclusion.

Solution:

$$Q_1 = L_1 + \frac{N/4 - C_f}{f}$$

$$\frac{N}{4} = \frac{100}{4} = 25$$

Quartile class = 600 - 800

$$Q_1 = 600 + \frac{25 - 24}{30} \times 200 = 606.67$$

Frequency distribution of expenditure of college students on fast food

Expenditure (in Rs.)	No. of students (f)	C.F.
200 - 400	5	5
400 - 600	20	24
600 - 800	30	54
800 - 1000	26	80
1000 - 1200	20	100

Thus, 75% of students spend above Rs. 606.67 on eating out at fast food joints.

4.4.3 Average Deviation (AD) or Mean Absolute Deviation (MAD)

Average deviation or mean absolute deviation is the average amount of variation of the items in a distribution from the average (mean or the median or the mode) ignoring the signs of these deviations. When mean is used as the average, we have the mean deviation about the mean, and when the median is used we have the mean deviation about the median. Similarly for mode it is mean deviation about mode.

Symbolically,

$$\text{A.D.} = \frac{1}{n} \sum |X_i - \bar{X}| \text{ for simple frequency distribution and about mean}$$

$$\text{A.D.} = \frac{1}{N} \sum f_i |X_i - \bar{X}| \text{ for grouped frequency distribution and about mean}$$

This measure is an improvement over the previous two measures in the sense that it considers all observations of a data set.

Example 4.38: A bank has designed a new process to serve its customers. This is meant to reduce the waiting time of customers. A random sample of 10 customers was taken after the new process was set up and the following waiting times were recorded. (in minutes)

3.5, 4, 3.7, 4.2, 5.0, 4.5, 3.8, 3.5, 3.2, 2.17

Find the average deviation.

Solution:

$$\text{A.D} = \frac{1}{n} \sum |x_i - \bar{x}|$$

$$\bar{x} = \frac{37.57}{10} = 3.76$$

Thus

$$\begin{aligned} \text{A.D} &= \frac{1}{10} \left[|3.5 - 3.76| + |4 - 3.76| + |3.7 - 3.76| + |4.2 - 3.76| + |5 - 3.76| + |4.5 - 3.76| + |3.8 - 3.76| \right. \\ &\quad \left. + |3.5 - 3.76| + |3.2 - 3.76| + |2.17 - 3.76| \right] \\ &= \frac{1}{10} [0.26 + 0.24 + 0.06 + 0.44 + 1.24 + 0.74 + 0.04 + 0.26 + 0.56 + 1.59] \\ &= \frac{5.43}{10} = 0.543 \end{aligned}$$

Example 4. 39: Find the average deviation about mean for the following distribution of demand for a book:

Frequency Distribution of Quantity Demanded

Quantity Demanded (in unit)	6	12	18	24	30	36	42
Frequency	4	7	10	18	12	7	2

Solution:

Calculation of mean deviation about mean

X_i	f_i	$f_i X_i$	$ x_i - \bar{x} $	$f_i x_i - \bar{x} $
6	4	24	17.6	70.4
12	7	84	11.6	81.2
18	10	180	5.6	56
24	18	432	.4	7.2
30	12	360	6.4	76.8
36	7	254	12.4	86.8
42	2	84	18.4	36.8
Total	60	1416		415.2

$$\text{The mean} = \frac{\sum_{i=1}^N f_i X_i}{N} = \frac{1416}{60} = 23.6$$

$$\text{The average deviation about mean} = \frac{1}{N} \sum f_i |X_i - \bar{X}| = \frac{415.2}{60} = 6.92$$

4.4.4 Standard Deviation (S. D.)

By far the most commonly used measure of dispersion is standard deviation. It is the square root of the arithmetic mean of the square of deviations of various values from their arithmetic mean. It is denoted by S.D or N.

The S.D. of a set of n observations X_1, X_2, \dots, X_n .

$$\sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \text{ for simple frequency distribution}$$

In simplified form this formula can be written as

$$\sigma_x = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2}$$

For grouped frequency distribution

$$\begin{aligned} \sigma_x^2 &= \frac{1}{N} \sum_{i=1}^n f_i (x_i - \bar{x})^2 \\ &= \frac{1}{N} \sum_{i=1}^n x_i^2 f_i - \left(\frac{1}{N} \sum_{i=1}^n x_i f_i\right)^2 \end{aligned}$$

$$\text{where } N = \sum_{i=1}^n f_i$$

The standard deviation is a measure of the degree of dispersion of the data from the mean value. It is a statistic that tells us how tightly all the various values are clustered around the mean in a set of data. A large standard deviation indicates that the data points are far from the mean and a small standard deviation indicates that they are clustered closely around the mean.

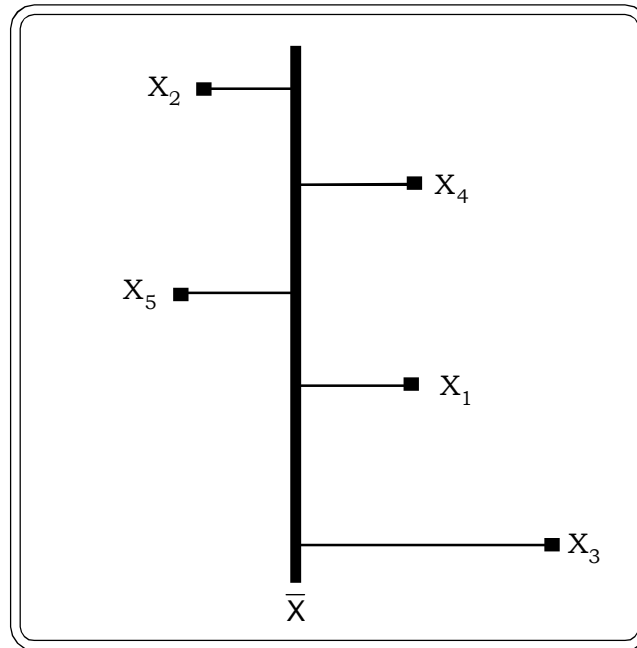


Figure 4.6

Deviations of observations from the mean

Standard deviation is rigidly defined and based on all observations. It is amenable to further algebraic treatment. It is not affected by sampling fluctuations and is less erratic. However, the problem with standard deviation is that it is difficult to understand and calculate and it gives greater weight to extreme values.

Variance

The term variance was used to describe the square of the standard deviation by R.A. Fisher in 1913. The concept of variance is of great importance in advanced work where it is possible to split the total into several parts, each attributable to one of the factors causing variations in their original series. It is the square of the standard deviation, as defined below, in case of a grouped frequency distribution.

$$\text{Variance} = \sigma^2 = \frac{1}{N} \sum_{i=1}^n f_i (X_i - \bar{X})^2$$

and in case of an ungrouped distribution,

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^n (X_i - \bar{X})^2$$

However, many authors used the above mentioned formula to calculate the standard deviation and variance from the entire population. Given only a sample of values x_1, \dots, x_n from some larger population, they define the sample standard deviation by

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

The reason for this definition is that s^2 is an unbiased estimator for the variance σ^2 of the underlying population. However, s is not an unbiased estimator for the standard deviation σ ; it tends to underestimate the population standard deviation. The concept of unbiasedness will be discussed in later chapters. In simple words an unbiased estimator is equally likely to assume values above the estimator and below the estimator.

Although the variance and standard deviation are equally valid measures of variability, the standard deviation is by far the more easily visualized and intuitively comprehended, because it is the one that is expressed in the same units of measurement as the original values of X_i of which the distribution is composed. When we calculate the standard deviation of our distribution, the resulting value of s also refers to the scale of the data. The variance, on the other hand, would refer to square units- square percentages, square dollars, square inches and so on, which do not readily lend themselves to graphic representation nor intuitive understanding.

4.4.4.1 Important Properties of SD

(i) Standard Deviation is independent of change of origin.

Proof:

$$\text{By definition } \sigma_x = \sqrt{\frac{1}{n} \sum (X_i - \bar{X})^2} \quad \dots 1$$

$$\text{Now } Y_i = X_i - A$$

$$\text{Or } X_i = A + Y_i \quad \dots 2$$

Now from the property of mean, for the present case, it can be written that

$$\bar{X} = A + \bar{Y} \quad \dots 3$$

$$X_i - \bar{X} = Y_i - \bar{Y}$$

$$\begin{aligned} \text{Thus } \sigma_x^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2 \\ &= \sigma_y^2 \end{aligned}$$

If $Y_i = X_i - A$ then $\sigma_x = \sigma_y$

- (ii) If $Y_i = \frac{X_i - A}{d}$ then $\sigma_x = d\sigma_y$ i.e. Standard Deviation is independent of change in origin but dependent of change in scale.

Proof:

$$\boxed{\text{If } Y_i = \frac{X_i - A}{d} \text{ then}}$$

$$X_i = A + dY_i \quad \dots (1)$$

$$\frac{1}{N} \sum_{i=1}^n X_i = A + d \frac{1}{N} \sum_{i=1}^n Y_i$$

$$\Rightarrow \bar{X} = A + d\bar{Y} \quad \dots (2)$$

$$(1) - (2)$$

$$\Rightarrow X_i - \bar{X} = d(Y_i - \bar{Y}) \quad \dots (3)$$

Now,

$$\sigma_x^2 = \frac{1}{N} \sum_{i=1}^n f_i (x_i - \bar{x})^2$$

$$= \frac{1}{N} \sum_{i=1}^n f_i (Y_i - \bar{Y})^2 d^2, \quad \text{using (3)}$$

$$= \left\{ \frac{1}{N} \sum_{i=1}^n f_i (Y_i - \bar{Y})^2 \right\} d^2$$

$$= \sigma_y^2 d^2$$

$$\Rightarrow \sigma_x = d\sigma_y$$

Thus S.D. is independent of change of origin but not of scale.

Thus, the following formulae for S.D are mathematically equivalent to the original S.D. formula as mentioned above

$$\boxed{\sigma_x = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2} = \sqrt{\frac{\sum fX^2}{N} - \bar{X}^2} = \sqrt{\frac{\sum fY^2}{N} - \left(\frac{\sum fY}{N}\right)^2} \times d \text{ where } Y = \frac{X - A}{d}}$$

4.4.4.2 Calculation of Standard Deviation

Example 4.40: Calculate standard deviation of the following observations

54, 55, 61, 60, 51, 59, 62, 52, 54, 49

Solution:

X	x - \bar{x}	(x - \bar{x})²
54	- 1.7	2.89
55	- 0.7	0.49
61	5.3	28.09
60	4.3	18.49
51	- 4.7	22.09
59	3.3	10.89
62	6.3	39.69
52	-3.7	13.69
54	-1.7	2.89
49	-6.7	44.89
		$\sum (X - \bar{X})^2 = 184.1$

$$\bar{X} = \frac{\sum X}{n} = \frac{557}{10} = 55.7$$

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$$

$$= \sqrt{\frac{184.1}{10}} = 4.29$$

Example 4.41: An Airline company recorded the following data regarding flight delays over a period of one week

Flight delays over a week

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
2	1	3	1	4	2	3

Find the average no. of delayed flights in a week. Compute the standard deviation.

Solution:

$$\begin{aligned}\text{Average no. of delayed flights} &= \frac{16}{7} \\ &= 2.28 \\ &\cong 2\end{aligned}$$

$$\begin{aligned}\text{Variance} &= \frac{1}{n} \sum x^2 - \bar{x}^2 \\ &= \frac{44}{7} - 4 = 2.28\end{aligned}$$

$$\text{Standard Deviation} = 1.51$$

Example 4.42: The following table shows the data which relate to the profit of 100 companies. Compute the variance and standard deviation.

Data relating to profits of a company

Profit (Rs Lakhs)	No. of Companies
8 - 10	8
10 - 12	12
12 - 14	20
14 - 16	30
16 - 18	20
18 - 20	10

Solution:**Calculation of S.D. for grouped data**

Profit (Rs Lakhs)	Mid points (x)	f	$Y = \frac{X-13}{2}$	fY	fY ²
8 - 10	9	8	-3	-24	72
10 - 12	11	12	-2	-24	48
12 - 14	13	20	-1	-20	20
14 - 16	15	30	0	0	0
16 - 18	17	20	1	20	20
18 - 20	19	10	2	20	40
	N = 100		-3	-28	200

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum fY^2}{N} - \left(\frac{\sum fY}{N}\right)^2} \times d \\ &= \sqrt{\frac{200}{100} - \left(\frac{-28}{100}\right)^2} \times 2 \\ &= \sqrt{2 - .0784} \times 2 \\ &= 1.3862 \times 2 = 2.77\end{aligned}$$

and Variance = $(2.77)^2 = 7.67$

Example 4.43: A consumer rights group is checking the price of a drug sold at different drug stores across the country. The drug was purchased from 25 stores located across the country. Find the variation in the price of the drug.

Price variation of a drug in various stores

Prices (Rs.)	No. of stores
6 - 8	7
8 - 10	12
10 - 12	6

Solution:

Calculation of variation in the price of a drug

Prices (Rs.)	No. of stores	x	xf	x ² f
6 - 8	7	7	49	343
8 - 10	12	9	108	972
10 - 12	6	11	66	726
	25		223	2041

$$\bar{x} = \frac{223}{25} = \text{Rs. } 8.92$$

$$\begin{aligned}\text{Variance} &= \frac{2041}{25} - 79.57 \\ &= 81.64 - 79.57 \\ &= 2.07\end{aligned}$$

Average price of the drug = Rs.8.92

Standard deviation of prices = Rs. 1.44

Example 4.44: The music of a new movie has been launched in the market. It is believed to be doing very well in terms of sales. The no. of CD's sold, in a survey of 40 cities shows the following distribution.

Sale of C.D's

No. of CD's Sold	No. of cities
500 - 1000	5
1000 - 1500	10
1500 - 2000	18
2000 - 2500	7

Find the mean number of CD's sold and also the standard deviation.

Solution:

Calculation of mean and standard deviation of number of CD's sold

No. of CD's Sold	No. of cities	x	$u = \frac{x - 1250}{500}$	fu	fu ²
500 - 1000	5	750	-1	-5	5
1000 - 1500	10	1250	0	0	0
1500 - 2000	18	1750	1	18	18
2000 - 2500	7	2250	2	14	28
				27	51

$$\begin{aligned}
 \text{Mean no. of CD's sold} &= 1250 + \frac{\sum fu}{\sum f} \times 500 \\
 &= 1250 + \frac{27}{40} \times 500 \\
 &= 1587.5 \cong 1588
 \end{aligned}$$

Standard Deviation

$$\begin{aligned}
 \sigma &= \sqrt{\frac{1}{N} \sum fu^2 - \left(\frac{\sum fu}{N} \right)^2} \times d \\
 &= \sqrt{\frac{51}{40} - \left(\frac{27}{40} \right)^2} \times 500
 \end{aligned}$$

$$= \sqrt{1.27 - 0.456} \times 500$$

$$= 0.9024 \times 500 = 451.193$$

Standard Deviation of the CD's sold = 451.193

4.4.4.3 Combined Standard Deviation

If two sets containing n_1 and n_2 items having means \bar{X}_1 and \bar{X}_2 and standard deviations σ_1 and σ_2 respectively are taken together then,

Standard Deviation of the combined set is defined as

$$\sigma = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

where $d_1 = \bar{X}_1 - \bar{X}$ and $d_2 = \bar{X}_2 - \bar{X}$. \bar{X} is the combined mean of the data.

Example 4.45: The following data is related to clients obtained by insurance agents during a given period for two types of insurance policies, a child policy and a retirement policy.

Data for two insurance policies

	Child Policy	Retirement Policy
No. of agents	25	18
Average No. of clients booked	72	64
Variance of the distribution	8	6

Calculate the combined standard deviation.

Solution:

There are two samples each of size 25 and 18 respectively. In order to calculate the combined mean we also need to know the averages of both the samples. From the Table 4.49,

Average of the first sample is 72 i.e. $\bar{X}_1 = 72$

And average of the second sample is 64 i.e. $\bar{X}_2 = 64$

Then combined Mean $\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2} = \frac{25 \times 72 + 18 \times 64}{25 + 18} = 68.6$

Now $d_1 = \bar{X}_1 - \bar{X} = 72 - 68.6 = 3.4$

$d_2 = \bar{X}_2 - \bar{X} = 64 - 68.6 = -4.6$

Combined variance

$$\begin{aligned}\sigma &= \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}} \\ &= \sqrt{\frac{25(8 + 11.56) + 18(6 + 21.16)}{25 + 18}} \\ &= \sqrt{22.74} = 4.77\end{aligned}$$

4.5 RELATIVE MEASURES OF VARIATION

As we know mean deviation, standard deviation etc. are the absolute measures of dispersion that expresses variation in the same units as the original data. To compare the variations (dispersion) of two different series, relative measures of standard deviation must be calculated. Three most commonly used relative measures of dispersion are:

- Co – efficient of Variation
- Co – efficient of Quartile deviation
- Co – efficient of Mean deviation

4.5.1 Co-efficient Of Variation (C. V.)

Co-efficient of variation or the co-efficient of standard deviation is the most important relative measures of dispersion. Actually it relates standard deviation and the mean by expressing the standard deviation as a percentage of the mean. This is a unit free measure and is in percentage form. Thus, the formula is

$$CV = \frac{\sigma}{\bar{X}} \times 100$$

where CV= Co-efficient of Variation

σ = Standard Deviation of the distribution

\bar{X} = Mean of the distribution

The coefficient of variation represents the ratio of the standard deviation to the mean, and it is therefore a useful statistic to compare the degree of variation from one data series to another, even if the means are drastically different from each other. In the investing world, the coefficient of variation allows us to determine how much volatility (risk) we are assuming in comparison to the amount of return one can expect from an investment. In simple language, the lower the ratio of standard deviation to mean return, the better the risk-return tradeoff.

Example:4.46: From the statistical record of a particular colony it is found that the monthly average income is Rs 42000 and S.D. is Rs. 1200. The average electric bill paid by them is Rs. 2460. and S.D Rs 120. State which is the less variable in nature the montly average income or the average electric bill.

Solution:

The given data is:

Mean and S.D. of income and electric bill

	Income (Rs)	Electric bill (Rs)
Mean (\bar{X})	42000	2460
Standard Deviation (σ)	1200	120

As the mean in the two items differ widely, even with the same unit, S.D is not the appropriate measure of variability here. Co-efficient of Variation (C.V) will be more appropriate here.

$$\text{We know } CV = \frac{\sigma}{\bar{X}} \times 100$$

$$\text{C.V of Income} = \frac{1200}{42000} \times 100 = 2.9\%$$

$$\text{C.V of Electric bill} = \frac{120}{2460} \times 100 = 4.9\%$$

Since Coefficient of variation for income is smaller, income is less variable than the electric bill.

Example 4.47: A bulb manufacturer buys filaments from 2 suppliers X and Y respectively. Due to some quality problems, which the manufacturer suspects could be because of the filaments, he wants to test which vendor is supplying better quality filaments. He takes samples of bulbs fitted with filaments from both the suppliers and obtains the following results.

Filament data of two suppliers

Length of Life (in hours)	Number of Bulbs Tested	
	Supplier (X)	Supplier (Y)
800 - 1000	30	28
1000 - 1200	40	35
1200 - 1400	35	25

Solution:

We will do the comparison of filaments of the two vendors using the co-efficient of variation.

For Supplier X,

Calculation of variability of filaments for supplier X

Length of Life	No. of Bulbs Tested (f)	x	fx	x ² f
800 - 1000	30	700	21000	14700000
1000 - 1200	40	1100	44000	48400000
1200 - 1400	35	1300	45500	59150000
		110500		

$$\bar{x}_x = \frac{110500}{105} = 1052.38 \text{ hours}$$

$$\sigma_x^2 = 1164285.7 - 1107503.7 = 56782$$

$$\sigma_x = 238.29 \text{ hours}$$

$$CV(x) = \frac{\sigma_x}{\bar{x}_x} = 0.2264$$

For Supplier Y

Calculation of variability of filaments of supplier Y

Length of Life	No. of Bulbs Tested (f)	x	fx	x ² f
800 - 1000	28	700	19600	13720000
1000 - 1200	35	1100	38500	42350000
1200 - 1400	25	1300	32500	42250000
	88			98320000

$$\bar{x}_y = \frac{90600}{88} = 1029.5455 \text{ hours}$$

$$\begin{aligned} \sigma_y^2 &= \frac{98320000}{88} - 1059963.9 \\ &= 1117272.7 - 1059963.9 = 57308.827 \end{aligned}$$

$$\sigma_y = 239.39 \text{ hours}$$

$$CV(y) = \frac{\sigma_y}{\bar{x}_y} = 0.2325$$

Comparing the two C.V.'s, variability is less in supplier X. Therefore, the filaments supplied by X seem to be more consistent.

4.5.2 Coefficient of Quartile Deviation

$$\text{Coefficient of Quartile Deviation} = \frac{\text{Quartile Deviation}}{\text{Median}} \times 100$$

4.5.3 Coefficient of Mean Deviation

$$\text{Coefficient of Mean Deviation} = \frac{\text{Mean Deviation}}{\text{Mean or Median}} \times 100$$

Here, if in the numerator, mean deviation is taken about the mean then the denominator is taken as the mean and similarly for median.

4.6 SKEWNESS

The term *skew* comes from an old French word meaning to shun or avoid. It is the same linguistic ancestor from which we get the English words 'eschew,' which also means to shun or avoid, and 'askew,' which conveys the meaning of lopsided or tilted off to one side. A skewed distribution is therefore one that is askew, lopsided, tending to shun or avoid one or the other of the extremes of the range within which it falls.

Examine the first two graphs in Figure 4.7 (repeated below) and you will see that the distributions of exam scores for Sections A and B are both conspicuously skewed, though in opposite directions. In Section A the exam scores tend to cluster toward the higher end of the range and taper off toward the lower end, whereas in Section B they tend to cluster toward the lower end of the range and taper off toward the higher end. In general, a distribution that is lopsidedly heavy at the higher end of the range and light at the lower end (e.g., Section A) is described as a *negatively skewed* distribution, while one that is heavy at the lower end and light at the higher end (e.g., Section B) is spoken of as a *positively skewed* distribution. Or to put it in pictorial terms, a negatively skewed distribution is one whose elongated tail extends to the left (the low or "negative") end of the range, while a positively skewed distribution is one whose elongated tail extends to the right (the high or "positive") end of the range.

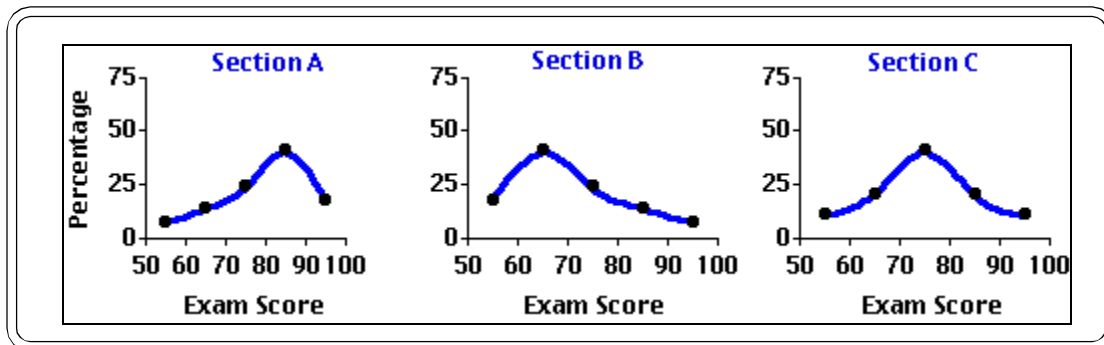


Fig. 4.7

Skewness in distribution of exam scores

An unskewed distribution, on the other hand, is one that is not lopsided in either the one direction or the other; typically it has two tails that trail off in both directions symmetrically. An example of an unskewed distribution is the third graph in Figure 4.7 showing the symmetrical distribution of exam scores in Section C.

A measure of the skewness of a distribution is given by

$$\frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

This measure is known as Karl Pearson's coefficient of skewness and lies between -3 and $+3$.

The skewness of a distribution can also be measured by comparing the mean median and mode of a distribution.

A distribution is said to be symmetric if $\text{mean} = \text{median} = \text{mode}$.

A distribution is positively skewed if $\text{mean} > \text{median} > \text{mode}$.

A distribution is negatively skewed if $\text{mean} < \text{median} < \text{mode}$.

Example 4.48 The mean, median and variance of a distribution is given by 45, 42 and 25 respectively. Calculate Karl Pearson's co-efficient of skewness.

Solution:

$$s_k = \frac{3(45 - 42)}{5} = 1.8$$

Karl Pearson's co efficient of skewness = 1.8

4.7 KURTOSIS

Kurtosis (from a Greek word meaning 'curvature' or 'convex') refers to whether the shape of a distribution is relatively short and flat, or tall and slender, or somewhere in-between those two extremes. If it is short and flat, like the distribution shown for Section D, it is described as *platykurtic* (flat-curved); if it is tall and slender, like the distribution shown for Section E, it is spoken of as *leptokurtic* (slender-curved); and if it is neither the one extreme nor the other, like the distribution shown for Section F, it is described as *mesokurtic* (medium-curved). In a platykurtic distribution the individual measures are spread out fairly uniformly across their range, whereas in a leptokurtic distribution they tend to cluster compactly at some particular point in the range.

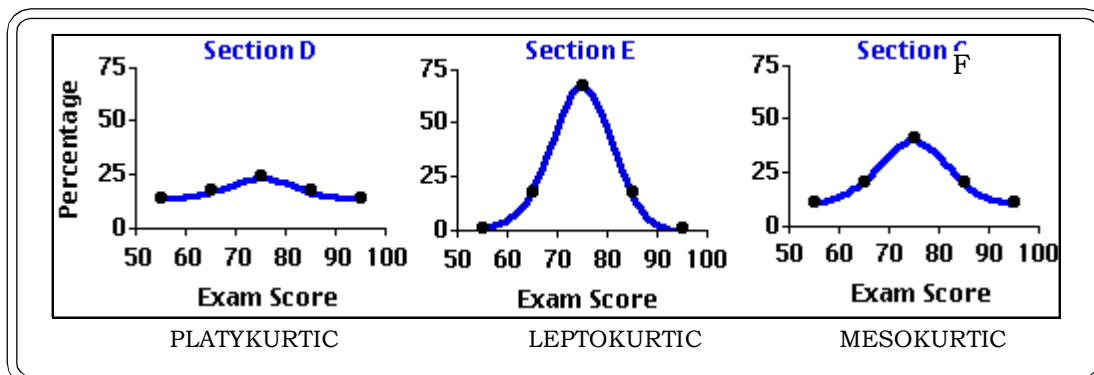


Fig. 4.8

Kurtosis

Thus, the exam scores for Section D are distributed fairly uniformly among the five intervals, with only a slight tendency to cluster and form a peak in the vicinity of the third interval (70 to 79.99...). The scores for Section E, on the other hand, have a quite pronounced tendency to cluster in the third interval, with only a few scores falling outside this interval. In the mesokurtic distribution illustrated by Section C the clustering is more moderate than in the leptokurtic distribution, and the curve as it falls away from the peak is more tapering than in the platykurtic distribution. A class containing approximately equal numbers of students whose mastery of the subject matter is very strong, quite strong, moderately strong, not so strong, and so on, would tend to produce the platykurtic distribution shown for Section D, while a class containing a large majority of students who all have approximately the same level of mastery would tend to produce a leptokurtic distribution of the kind shown for Section E.

4.8 CASELETS

Caselet 1: The president of the group of publication of the magazine 'Today's India' is interested in studying the sales of the magazine in 25 cities across different zones of India. The data sent in from 25 cities has been compiled at the headquarters in the form of a frequency distribution. The figures of sales frequencies of the magazine that confront the president are as follows

Table 4.2

Sales (000)	Frequency
0 - 5000	2
5000 - 10000	6
10000 - 15000	10
15000 - 20000	5
20000 - 25000	2

The president wants to study this distribution and look for answers for the following questions

- What is the overall average sales figure of the magazine all over India?
- How many cities have sales exceeding the average and below the average?
- How much variability is there in terms of sales in different cities surveyed?

Caselet 2: An electronic company has two suppliers for wires for their printed circuit boards (PCB). Lately there have been reports of quality problems from the production department about the breaking strength of the wires. To test the performance of the individual suppliers, two samples of size 5 each were taken from two lots—one of supplier A and one of supplier B. The results are

A: 290 304 296 312 325

B: 275 325 350 305 375

The company wants to find how consistent both the suppliers are.

4.9 EXCEL GUIDE

Calculation of Mean

We can use the Excel spreadsheet program to assist us in many calculations, including calculation of the mean.

Let us say that we have the quiz scores for nine students in a Quantitative techniques class for their first four quizzes. The data for these nine pupils could appear in a spreadsheet as follows:

	A	B	C	D	E
1	Quiz Scores for Students in Mathematics Class				
2	Name	Quiz 1	Quiz 2	Quiz 3	Quiz 4
3	Aaron	19	18	18	16
4	Betty	12	13	19	17
5	Daniel	18	19	17	19
6	James	17	16	18	18
7	Joseph	19	18	17	20
8	Karen	13	14	16	19
9	Lois	15	14	19	18
10	Mary	17	18	17	17
11	Peter	14	13	18	16
12					
13					
14					

Using the of Excel starting with the worksheet containing the quiz scores we can find the mean or average for each quiz by proceeding as follows:

- (1) Click on the cell **B13** and click on which is the paste function icon.
- (2) In the **Paste Function** window which appears select **Statistical** under **Function Category:** and **Average** under **Function name:**, and click **OK**.
- (3) In the **Number 1** box that appears enter **B3:B11** and click **OK**. The mean for Quiz 1 (based on the scores in cells B3 through B11) appears in cell **B13** and has the value **16**.
- (4) To format this value to two decimal places, select **Cells** from the **Format** menu. In the **Format Cells** window that appears click on the **Numbers** tab and click on **Numbers** under **Category:**.
- (5) Adjust the **Decimal places:** box so that is shows **2**. You can adjust the number in the box by clicking on the up or down arrows to the right of the box. Click **OK** and observe that the mean for Quiz 1 appears as **16.00**
- (6) Copy this same formula to the other three quizzes by clicking on cell **B13** and dragging over to cell **E13**. Select **Fill** from the **Edit** menu and then with the mouse button depressed slide to the right and select **Right**.

The completed worksheet shows the means for all four quizzes and should look something like the following:

	A	B	C	D	E
1	Quiz Scores for Students in Mathematics Class				
2	Name	Quiz 1	Quiz 2	Quiz 3	Quiz 4
3	Aaron	19	18	18	16
4	Betty	12	13	19	17
5	Daniel	18	19	17	19
6	James	17	16	18	18
7	Joseph	19	18	17	20
8	Karen	13	14	16	19
9	Lois	15	14	19	18
10	Mary	17	18	17	17
11	Peter	14	13	18	16
12					
13	Mean	16.00	15.89	17.67	17.78
14					

Standard Deviation

In Microsoft Excel, type the following code into the cell where you want the Standard Deviation result, using the “unbiased,” or “n – 1” method:

= STDEV (A1 : An) (substitute the cell name of the first value in your dataset for A1, and the cell name of the last value for An.)

or use = STDEVP(A1 : An) if you want to use the “biased” or “n” method

4.10 EXERCISES

- 4.1 What do you understand by the concept of central tendency?
- 4.2 What are the various measures of central tendency?
- 4.3 Explain the important properties of the arithmetic mean. Why is it considered to be the best measure of central tendency?
- 4.4 Give an illustration of when you could use a geometric mean instead of an arithmetic mean?
- 4.5 Explain the concepts of quartiles, deciles & percentiles. How would you locate them graphically?
- 4.6 When would you use the mode? Give few examples.
- 4.7 Explain the concept of variation in data?
- 4.8 What are the various measures of variation or dispersion?
- 4.9 State the important properties of dispersion.
- 4.10 What are relative measures of dispersion? How are they different from the absolute measures of dispersion?
- 4.11 Explain the concepts of skewness and kurtosis with examples.
- 4.12 Calculate the mean, standard deviation and variance for the following data.

No. of defects per item	Frequency
0 - 5	18
5 - 10	32
10 - 15	50
15 - 20	75
20 - 25	125
25 - 30	150
30 - 35	100
35 - 40	90
40 - 45	80
45 - 50	50

- 4.13 A personal manager is interested to know the average length of stay of its employees in the organization. The following data shows the record of organization:

Years	No. of employee
1 - 2	15
2 - 3	20
3 - 4	15
4 - 5	12
5 - 6	5

Calculate the average numbers of years for which an employee works in the company before leaving it.

- 4.14 Calculate geometric mean of the following price relatives of food items

Commodity	Price Relatives
Rice	208
Wheat	307
Pulses	156
Sugar	140
Tea	200

4.15 Given below is the wage distribution of 100 workers in a factory

Wages (Rs)	No. of workers
Below 1000	3
1000 - 1200	5
1200 - 1400	12
1400 - 1600	23
1600 - 1800	31
1800 - 2000	10
2000 - 2200	8
2200 - 2400	5
2400 and above	3

Determine graphically the values of Q_2 , Q_3 , D_{55} and P_{18} and verify the results by the corresponding mathematical formula.

4.16 From a survey of 60 chemical industries following data are collected.

Level of profit(Rs Lakhs)	No. of companies
15 - 20	20
20 - 25	14
25 - 30	13
30 - 35	8
35 - 40	5

Calculate the variance for the distribution.

4.17 A company has three establishments A_1 , A_2 and A_3 in three cities. Analysis of the monthly salaries paid to the employees in the three establishments is given below:

	A_1	A_2	A_3
No. of employees	50	30	20
Average monthly salary	3050	4000	4200

Find the average monthly salary of all the 100 employees in the company.

- 4.18 A factory produces two types of electric lamp. The following are the life expectancy of the lamps. Compare the variability of the life of the two types of electric lamps using the coefficient of variation.

Life length (Hours)	Type 1	Type 2
600 - 700	4	3
700 - 800	8	5
800 - 900	9	7
900 - 1000	7	12
1000 - 1100	3	4

- 4.19 From the data given below draw and find out the median and the first quartile.

Marks	Frequency
0 - 10	9
10 - 20	15
20 - 30	23
30 - 40	18
40 - 50	10
50 - 60	5

- 4.20 Fluctuation of the daily sales of two products X and Y are given below. Find out which of the two products shows greater fluctuation in sales.

Sales of Product X	520	524	522	525	518	517	523	526
Sales of Product Y	1152	1134	1133	1146	1132	1134	1149	1130

- 4.21 The percent annual growth rates of output of a particular factory in 5 years from 2001 to 2005 are 5, 7.5, 2.5, 5 and 10 respectively. What is the annual compound rate of growth for the entire period of 5 years?

4.22 The table below gives the amount of petrol sold at a petrol pump on a given day.

Amount of Petrol (in liters)	No. of cars
0 - 5	15
5 - 10	80
10 - 15	110
15 - 20	65
	270

(i) Find the average amount of petrol sold.

(ii) Find the modal class and the mode.

(iii) Calculate the variance and the standard deviation.

4.23 In a class of 50 students, 25 girls had a average weight of 532kgs. and 25 boys had an average of 60kgs. Find the average weight of the students of this class.

4.24 Find the mode & median for the following data:

Daily Sales (in Thousands)	No. of Companies
10 - 20	15
20 - 30	23
30 - 40	27
40 - 50	20
50 - 60	35
60 - 70	25
70 - 80	5

4.25 Three workers are doing a certain task. The first worker takes 3 minutes to complete the task, the second worker takes 5 minutes to finish the task and the third worker can complete the task in 4 minutes. Find the average time taken by the workers to complete the task.

4.26 A call center compiled the following data regarding the number of calls received.

No. of Calls (per day)	No. of Days
100 - 200	3
200 - 300	11
300 - 400	13
400 - 500	12
500 - 600	7

Find the mean, median, mode and variance of the distribution of calls.

4.27 Given that the median of the following distribution is 46. Find the missing frequencies.

Class Interval	No. of Days
10 - 20	12
20 - 30	30
30 - 40	X
40 - 50	40
50 - 60	Y
60 - 70	25

4.28 The following is the average amount of dollars each major airline spends per passenger on food:

American	7.41
United	7.24
Northwest	5.15
TWA	5.09
Delta	4.61
Continental	2.77
US Air	2.68
American West	2.00

What are the mean and median cost per passenger? Which would be the better figure to use for a new airline in developing its business plan? **(MBA, DU, June 2003)**

4.29 (a) A machine is assumed to depreciate 44% in value in the first year, 25% in the second year and 10% per annum for the next three years, each percentage being calculated on the diminishing value. What is the average percentage depreciation for the five years?

(MBA, Vikram Univ., 2001)

(b) Mr. A spends Rs.1000 for apples costing Rs.25 per kilogram and another Rs.1000 for apples costing Rs.20 per kilogram. What is the average price of apples per kilogram?

(MBA, Vikram Univ., 2001)

4.30 Find the missing frequencies in the following distribution if N is 100 and median 30:

Marks	No. of Students
0 - 10	10
10 - 20	15
20 - 30	?
30 - 40	30
40 - 50	10
50 - 60	8

4.31 Calculate the mean and median for the following data:

(MBA, Madurai – Kamraj Univ., Nov., 2003)

Central Wages (in Rs.)	No. of Wage Earners
15	3
20	25
25	19
30	16
35	4
40	5
45	6

4.32 Lives of two models of refrigerators in a recent survey are:

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

What is the average life of each of these refrigerators? Which model has greater uniformity?

(MBA, Bharthidasan Univ., 2001; IAS 2002; CBSE, 2002)

4.33 The life of two types of tyres in a sample survey is given below:

Life (in Km.)	Type A	Type B
5000 - 10000	18	15
10000 - 15000	22	24
15000 - 20000	26	30
20000 - 25000	25	18
25000 - 30000	9	13

(a) Which of the two types of tyre give a higher average life?

(b) If prices are same for both the types which would you prefer and why? *(MBA, DU, 1999)*

4.34 Name of the various measures of dispersion. How would you compare the performance of two companies, which reported profits for last five years as follows: *(MBA, DU, 2000)*

Company I	Company I
4.0	7.3
4.1	-3.7
4.3	8.4
4.0	-2.5
4.1	11.0

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

Age Groups (in yrs.)	Amount of Scholarship per month (Rs.)
5 - 7	300
8 - 10	400
11 - 13	500
14 - 16	600
17 - 19	700

The ages (years) of 30 school going children are noted as 11, 8, 10, 5, 7, 12, 7, 17, 5, 13, 9, 8, 10, 15, 7, 12, 6, 7, 8, 11, 14, 18, 6, 13, 9, 10, 6, 15, 3, 5 years respectively. Calculate mean and standard deviation of monthly scholarship. Find out the total monthly scholarship amount being paid to the students. *(MBA, IGNOU, 2002)*

4.36 Given below are daily wages in rupees of 60 workers in a factory manufacturing plastic products:

23	48	51	64	72	82	56	33	50	42
35	88	77	65	39	52	48	64	49	57
41	73	62	49	32	54	67	46	55	50
82	44	75	56	51	63	59	69	53	42
75	85	68	55	52	45	42	57	20	57
46	51	20	16	62	46	54	40	55	71

(a) Form a frequency distribution, taking the lowest class – interval as 10 – 20.

(b) Calculate the Standard Deviation and Coefficient of Variation of this distribution.

(MBA, HPU, 2002)

4.37 The following data give the number of finished articles turned out per day by different number of workers in a factory:

No. of articles	No. of workers
18	3
19	7
20	11
21	14
22	18
23	17
24	13
25	8
26	5
27	4

Find the mean, standard deviation and coefficient of variation of daily output of finished articles.

(MBA, Kurukshetra Univ., 2001)

4.38 The prices of a Tea Company shares 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 the prices of shares. In which market are the shares prices stable. **(MBA, HPU, 2002)**

4.39 You are given the data pertaining to kilowatt hours of electricity consumed by 100 persons in Delhi.

Consumption (K. watt hours)	No. of users
0 but less than 10	6
10 but less than 20	25
20 but less than 30	36
30 but less than 40	20
40 but less than 50	13

Calculate (i) standard deviation and (ii) the range within which middle 50% of the consumers fail. **(MBA, DU, 1996)**

4.40 The following data gives the consumption of electricity in terms of number of units consumed. Calculate the quartile deviation.

No. of Units	No. of Consumers (in thousands)
200 - 400	10
400 - 600	15
600 - 800	23
800 - 1000	17
1000 - 1200	9



5

Probability and Probability Distributions



Structure

- 5.1 Introduction
- 5.2 Set Theory
 - 5.2.1 Definition of a Set
 - 5.2.2 Types of Sets
 - 5.2.3 Pictorial Representation of Set Theory - Venn Diagrams
 - 5.2.4 Properties of Set Operation
- 5.3 Counting Rules
 - 5.3.1 Permutations
 - 5.3.2 Combinations
- 5.4 Some Important Terms in Probability
 - 5.4.1 A Random Experiment
 - 5.4.2 Sample Space
 - 5.4.3 Trial
 - 5.4.4 Event
 - 5.4.4.1 Rules of Event Operations
 - 5.4.4.2 Types of Events
- 5.5 Various Definitions of Probability
 - 5.5.1 The Theoretical Definition of Probability
 - 5.5.2 The Classical Theory of Probability
 - 5.5.3 Relative Frequency or Empirical Approach
 - 5.5.4 Axiomatic Approach
 - 5.5.5 Subjective Approach

- 5.6 Laws and Theorems of Probability**
 - 5.6.1 Additive law**
 - 5.6.2 Conditional Probability and Multiplication Law of Probability**
 - 5.6.3 Theory of Independence**
 - 5.6.4 Pair-wise and Mutual Independence**
 - 5.6.5 The Theorem of Total Probability and the Baye's Theorem**
- 5.7 Probability Distribution of a Random Variable**
 - 5.7.1 Random Variables**
 - 5.7.2 Discrete Random Variable**
 - 5.7.3 Continuous Random Variable**
 - 5.7.4 Probability Distribution of a Random Variable**
- 5.8 Discrete Probability Distributions**
 - 5.8.1 Expected value and Variance of a Discrete Probability Distribution**
 - 5.8.2 Binomial Distribution**
 - 5.8.2.1 Mean of Binomial Distribution**
 - 5.8.2.2 Variance of Binomial Distribution**
 - 5.8.2.3 Mode of the Binomial Distribution**
 - 5.8.2.4 Fitting of Binomial Distribution**
 - 5.8.3 Poisson Distribution**
 - 5.8.3.1 Mean of Poisson Distribution**
 - 5.8.3.2 Variance of Poisson Distribution**
 - 5.8.3.3 Mode of the Poisson Distribution**
 - 5.8.3.4 Poisson Approximation to Binomial**
 - 5.8.3.5 An Application of the Poisson Distribution**
 - 5.8.3.6 Fitting of Poisson Distribution**
- 5.9 Continuous Probability Distribution**
 - 5.9.1 Normal Distribution**
 - 5.9.2 Characteristics of Normal Distribution**
 - 5.9.3 Normal as an Approximation to Binomial Distribution**
- 5.10 Caselets**
- 5.11 Excel Guide**
- 5.12 Exercises**

5.1 INTRODUCTION

In everyday life we constantly make statements which are probabilistic in nature and involves an element of uncertainty. For example, we may wonder one cloudy morning whether it is going to rain and evaluate what is the chance that it is going to rain. More examples would be wondering if our flight would be on time, or the possibility of getting a movie ticket for a popular movie and so on. All these events are random in nature and probability theory includes studying such events to make the best possible decision in the face of uncertainties. Thus, probability is a study of random events in an attempt to rationalize randomness. It is an essential tool of analysis of modern business and economic problems.

The history of probability theory goes back to the middle of the seventeenth century. Two contemporary French mathematicians Blaise Pascal (1623-1662) and Pierre Fermat (1601-1665) are credited with most of the developments of probability theory.

5.2 SET THEORY

Since set theory, developed by the German mathematician G. Cantor plays a fundamental role in developing probability, we discuss briefly some basic concepts related to set theory before proceeding further.

5.2.1 Definition of a set

Set:

A set is a well-defined collection of objects. For example, if A is a set of 5 natural numbers, then the set A can be symbolized as

$$A = \{1, 2, 3, 4, 5\}$$

Element of a Set: If A is a set and z_i is a member of it, then z_i is called element of Set A and is denoted by.

$$z_i \in A \quad \text{where } \in \text{ - belongs to}$$

In the above example, 1, 2, 3, 4 and 5 are elements of A.

5.2.2 Types of set

- (i) **Universal Set:** The universal set is the set of all sets. And all sets are sub sets of it.
- (ii) **Empty Set:** Empty sets contain no elements and are usually denoted by ϕ . An Empty set is also known as a null set.
- (iii) **Subset:** Any sub collection of objects from a set A is called a subset of A. For example if

$$A = \{a, b, c, d, e, f\} \text{ is a set}$$

$$\text{then } A\phi = \{a, c, e\} \text{ is a sub set of A}$$

- (iv) **Intersection of Set:** Intersection of two sets is a new set consisting of the common elements of the two original sets. For example, let

$$A = \{1, 2, 3, 4, 5\}$$

$$\text{and } B = \{4, 5, 6\}$$

Then, intersection of A and B denoted symbolically by $A \cap B$ can be a new set with the following common elements

$$A \cap B = \{4, 5\}$$

(v) Union Set: Union set is the set of elements that are in at least one of the two sets. For example, as before, let

$$A = \{1, 2, 3, 4, 5\}$$

$$B = \{4, 5, 6\}$$

The union of A and B i.e. $A \cup B$ is a set of 6 elements:

$$A \cup B = \{1, 2, 3, 4, 5, 6\}$$

(vi) Difference of Set: The difference of two sets say A and B consists of elements of A minus the elements of B. For example:

If $A = \{1, 2, 3, 4, 5\}$

$$B = \{1, 2, 3\}$$

$$A - B = \{4, 5\}$$

(vii) Equal Set: Two sets are said to be equal if all the elements of one set are also the element of the other set. Equal sets are denoted by:

$$A = B$$

(viii) Disjoint Set: If two sets never intersect each other, they are known as disjoint set.

If $A = \{1, 2, 3, 4, 5\}$

$$B = \{6, 7, 8\}$$

A and B are disjoint. They have no common elements.

Thus, some standard symbols used in set theory are given in the following table:

Table 5.1
Symbols used in set theory

Symbol	Name
ϕ	Null set
\in	Element of
\notin	Not an element of
\subseteq	Contained in
\supseteq	Contains
$\not\subseteq$	Not contained in
\cap	Intersection
\cup	Union

5.2.3 Pictorial Representation of Sets – Venn Diagrams

The concepts of sets and their relations can be easily visualized pictorially by diagrams known as Venn diagrams. Venn Diagrams have been developed by the English mathematician John Venn (1834 – 1923). They are very useful in depicting relationships between sets.

The following figures depict the pictorial representation of the types of sets.

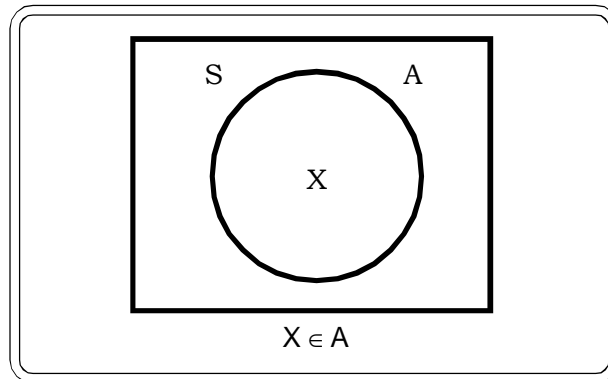


Figure 5.1

Element of a set

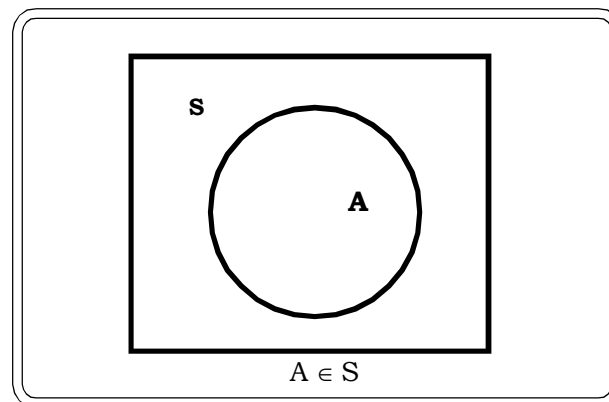


Figure 5.2

Subset

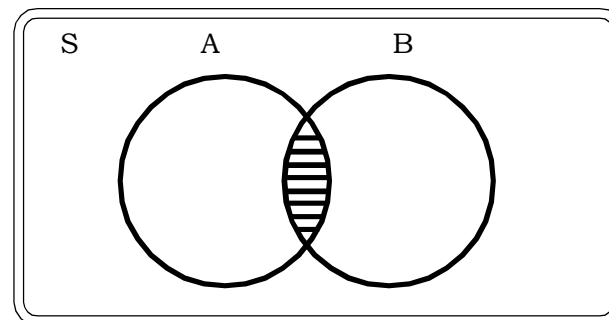


Figure 5.3

Intersection of two sets A & B

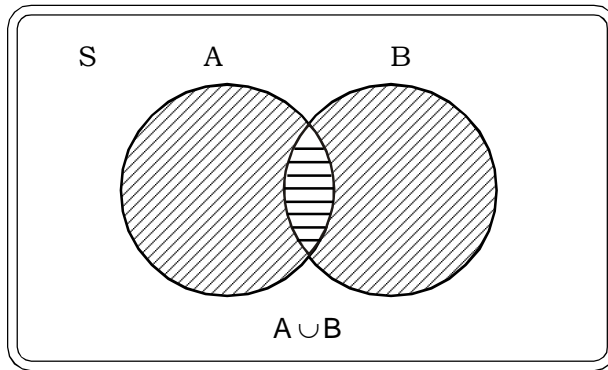
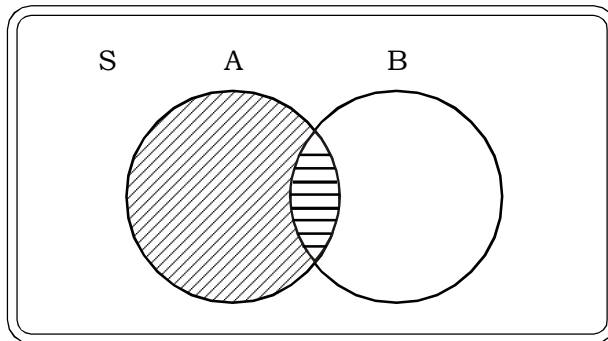


Figure 5.4

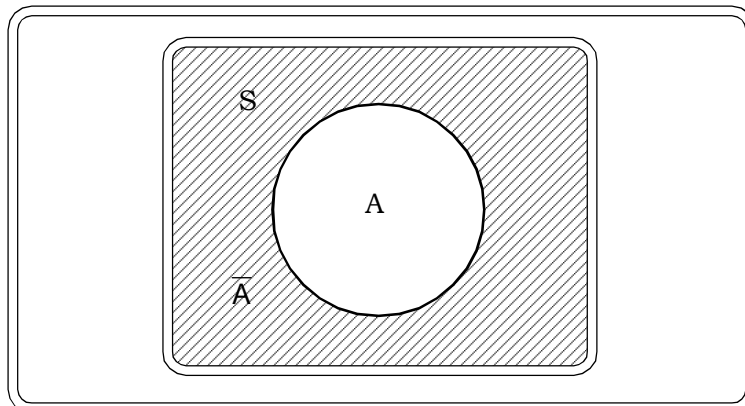
Union of Sets



$A - B$

Figure 5.5

Difference of Two Sets



\bar{A} or A^c

Figure 5.6

Complement of a Set

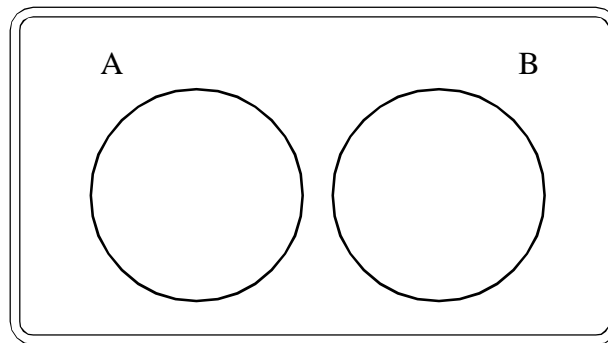


Figure 5.7

Disjoint Sets

The rectangular region in all the above figures represent the Universal set.

5.2.4 Properties of Set Operation**1. Identity Laws:**

$$A \cup \phi = A \quad \text{where } \phi \text{ null set}$$

$$A \cap S = A$$

2. Domination law:

$$A \cup S = S$$

$$A \cap \phi = \phi$$

3. Idempotent Laws:

$$A \cup A = A$$

$$A \cap A = A$$

4. Cumulative Laws:

$$A \cup B = B \cup A$$

$$A \cap B = B \cap A$$

5. Associative Laws:

$$(A \cup B) \cup C = A \cup (B \cup C)$$

$$(A \cap B) \cap C = A \cap (B \cap C)$$

6. Distributive Laws:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

7. De Morgan's Laws:

$$A - (B \cup C) = (A - B) \cap (A - C)$$

$$A - (B \cap C) = (A - B) \cup (A - C)$$

8. Law of Complement:

$$\overline{\overline{A}} = A$$

5.3 COUNTING RULES 

Counting rules like permutations and combinations are fundamental for calculation of various probabilities. In this section, we discuss rules of permutations and combinations with examples. These rules are useful in computing total number of outcomes of a random experiment and also favourable cases of occurrence of any event, which are fundamental to the classical definition of probability. (described in section 5.5)

5.3.1 Permutations

A permutation is an arrangement of a given number of objects in a particular order.

Different kinds of permutations possible are:

(i) Permutation of n distinct objects

The total number of permutation of n distinct objects is $n!$

Thus, permutation of n objects, all taken together is

$${}^n P_n = \frac{n!}{(n-n)!} = n!$$

For example, suppose n children are to be seated on n chairs. The first chair can be occupied by any one child in n ways. Similarly, the second chair can be occupied in $(n-1)$ ways and so on till the last chair can be occupied by the last child in one way. Thus, the total number of ways in which the n children can be seated on n chairs is:

$${}^n P_n = n(n-1)(n-2) \dots 3.2.1 = n!$$

(ii) Permutations of n objects taking r at a time.

Now, suppose the n children are to be seated on n chairs, where $r \leq n$.

The total no of arrangements or permutations of n children taken r at a time is given by

$$\begin{aligned} {}^n P_r &= \frac{n!}{(n-r)!} \\ &= \frac{n(n-1)(n-2) \dots [(n-(r-1)](n-r)!}{(n-r)!} \\ &= n(n-1)(n-2) \dots (n-r+1) \end{aligned}$$

(iii) Permutation of n objects taken r at a time, when any object may be repeated any number of times.

Suppose each of the r chairs are to be filled with the n children .

Then, the total no of arrangements or permutations possible are

$$\underbrace{n \times n \times n \dots \times n}_{r \text{ times}} = n^r \text{ ways}$$

(iv) Permutations of n objects in a circular order

The number of permutations of n objects in a circular order is given by

$$\frac{{}^n P_n}{n} = \frac{n!}{n} = (n - 1)!$$

(v) Permutations of n objects of different kinds

Suppose there are n objects such that n_1 are of one kind, n_2 are of another kind, and n_k of a kind. then, the total ways in which these n objects (where $n = n_1 + n_2 + \dots + n_k$) can be arranged or the no of permutations of the n objects are

$$\frac{n!}{n_1!n_2!\dots n_k!}$$

Example 5.1: A consumer is asked to rank 4 different kinds of ice-creams according to his or her preference. How many ranking are possible in all?

Solution: Total no of possible rankings = ${}^4 P_4$
 $= 4! = 24$

Example 5.2:

- 12 seats are available on a bus and there are 10 people. In how many possible ways can the 10 people be arranged?
- In a beauty contest with 20 participants, three prizes are to be awarded viz: one for talent round, one for national costume and one for evening wear. In how many ways can the prizes be awarded?
- In how many ways can the letters of the word ABSENT arranged?
- In how many ways can the letters of the word PROBABILITY arranged?
- In how many ways can 10 children sit around in a circle for a game of passing the parcel?
- In how many ways can 2 boys and 2 girls be seated such that boys and girls occupy alternative positions?

Solution

- (a) The total number of ways in which 10 people can be arranged on 12 seats are

$${}^{12} P_{10} = \frac{12!}{10!} = 132 \text{ ways}$$

- (b) Each prize can be awarded in 20 ways as each prize can be given to the same contestant.

\therefore Total no of ways in which the three prizes can be awarded is

$$\begin{aligned} &= 20 \times 20 \times 20 \\ &= 20^3 \\ &= 8000 \text{ ways} \end{aligned}$$

- (c) There are 6 letters in the word ABSENT and the number of ways in which 6 letters can be arranged amongst themselves is

$$\begin{aligned}
 {}^6P_6 &= \frac{6!}{(6-6)!} = 6! \\
 &= 6 \times 5 \times 4 \times 3 \times 2 \times 1 \\
 &= 720 \text{ ways}
 \end{aligned}$$

- (d) The word PROBABILITY has 11 letters in all. Out of these, there are 2 B's and 2 I's.

Thus, the total no of permutations of the letters of the word PROBABILITY are

$$\begin{aligned}
 &= \frac{11!}{2!2!} \\
 &= \frac{11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 4 \cdot 3}{2} \\
 &= 1995840 \text{ ways}
 \end{aligned}$$

- (e) 10 children can sit around in a circle for a game of passing the parcel in

$$(10 - 1)! = 9! = 362880 \text{ ways.}$$

- (f) There are two ways in which 2 boys and 2 girls can be seated alternatively i.e.

B G B G and G B G B

The girls can rearrange themselves in $2!$ ways and the boys can rearrange themselves in $2!$ ways in each of the above arrangements

\therefore The reqd no of permutations or the total number of ways in which two boys and two girls can be arranged such that boys and girls occupy alternative positions is

$$= 2 \times 2! \cdot 2! = 8$$

Example 5.3: A manager at a book store has just received 3 childrens books, 2 classics and 4 management books. In how many ways can he arrange these books on a shelf if

- all the books are arranged at random?
- books of each category are arranged together?
- only the children's books are arranged together?
- management books and the rest of the books are arranged together?

Solution:

- (a) There are 9 books in all and these books can be arranged in

$$9! = 3,62,880 \text{ ways}$$

- (b) The children books can be arranged in $3!$ ways

The 2 classics books can be aranged in $2!$ ways

and the management books can be arranged in $4!$ ways

The 3 groups can be arranged among themselves in $3!$ ways

Thus, the total number of ways in which books of each category can be arranged together

$$= 3! \cdot 2! \cdot 4! \cdot 3!$$

$$= 1728 \text{ ways}$$

(c) Considering the childrens books as one book the total number of books = 7

Possible arrangements of 7 books = 7!

The children's books can be further arranged amongst themselves in 3! ways.

Thus, the total number of permtations possible

$$= 7! 3!$$

$$= 30240 \text{ ways}$$

(d) Considering the management books as one group and the rest of the books as another group, both the groups can be arranged in 2! ways. The management books can be arranged in 4! ways and the rest of the books can be arranged in 5! ways.

Therefore the total number of ways = 2! 4! 5!

$$= 5760 \text{ ways}$$

5.3.2 Combinations

When r objects are to be selected out of n objects without any consideration about the order of arrangement, then we get a combination. This is denoted by

$$\frac{{}^n P_r}{r!} = \frac{n!}{r!(n-r)!} \quad \frac{{}^n P_r}{r!} = \frac{n!}{r!(n-r)!} = {}^n C_r$$

$$\text{Result 1: } {}^n C_r = {}^n C_{n-r}$$

Result 2: The total number of combinations of n distinct objects taken 1, 2, 3,, n at a time is

$$\begin{aligned} & {}^n C_1 + {}^n C_2 + {}^n C_3 + \dots + {}^n C_n \\ &= 2^n - 1 \end{aligned}$$

Example 5.4:

- In how many ways can 3 balls be selected out of a total of 10 balls?
- Suppose a box contains 5 black balls and 6 red balls. In how many ways can 2 black ball and 3 red balls be drawn from the box? (when the balls are replaced)
- There are 10 pens packed in a box out of which two are defective. In how many ways can 4 pens be selected out of the 10 pens so as to include atleast two defectives.

Solution:

(a) The number of ways in which three balls can be selected out of 10 balls = ${}^{10}C_3$

$$\begin{aligned} &= \frac{10!}{3!7!} = \frac{10 \cdot 9 \cdot 8 \cdot}{3 \cdot 2} \\ &= 120 \text{ ways} \end{aligned}$$

(b) Out of 5 black balls, 2 balck balls can be drawn in 5C_2 ways.

Out of 6 red balls, 3 red balls can be drawn in 6C_3 ways,

Thus, no of ways in which 2 black balls and 3 red balls can be drawn from a box of 5 black and 6 red balls is ${}^5C_2 \times {}^6C_3$

(c) The no. of defectives may be one or two

(i) Thus, the no of ways in which one defective pen would be included

$$= {}^8C_3 \times {}^2C_1 = \frac{8 \cdot 7 \cdot 6}{3 \cdot 2} \times 2 = 112$$

(ii) The no of ways in which 2 defective pens are included

$$= {}^8C_2 \times {}^2C_2 = \frac{8 \cdot 7}{2} \times 1 = 28$$

The total number of ways of selecting 4 pens such that atleast two of them are defective.

$$= {}^8C_3 \times {}^2C_1 + {}^8C_2 \times {}^2C_2 = 112 + 28 = 140$$

5.4 SOME IMPORTANT TERMS IN PROBABILITY

5.4.1 A Random Experiment: A random experiment or a random phenomenon is a procedure that can result in many outcomes such that although all outcomes may be known it is not possible to predict the outcome associated with a single experiment. For example, tossing of a coin is a random experiment whose all possible outcomes are head or tail, but at each particular toss, we don't know if it will face head or tail.

5.4.2 Sample Space: The sample space of a random experiment is the set S that includes all possible outcomes of an experiment. It plays the role of the universal set when modelling an experiment. For example, in the tossing of a fair coin the set of all possible outcomes is {H, T}. Thus {H, T} is the sample space for this experiment. Also when a dice is tossed the sample space will be {1, 2, 3, 4, 5, 6}. Each element in the sample space is called a sample point.

A sample space with a finite number of outcomes is known as a finite sample space.

If e_1, e_2, \dots, e_n are the n outcomes of an experiment, then the sample space is written as

$$S = \{e_1, e_2, \dots, e_n\}$$

Example 5.5: Suppose a committee of 2 is to be selected from a group consisting of 5 people say Rahul (R), Siddharth (S), Vaibhav (V), Ankit (A) and Neha (N). Find the sample space.

Solution:

The sample space will be

$$S = \{RS, RV, RA, RN, SV, SA, SN, VA, VN, AN\}$$

Example 5.6: Suppose three children are born to a family. Denoting the birth of a daughter by f and the birth of a son by m , write down the sample space.

Solution:

The required sample space is

$$S = \{mff, mfm, mmm, fmm, fmf, fff, mmf, ffm\}$$

Example 5.7: A manufacturer purchases equipments from three vendor A, B, C. Let (1,2) denote the event that on two successive days, the first day order goes to vendor A and on the second day vendor B gets the order. Write the sample space.

Solution:

The 9 elements of the sample space are

$$S = \{(1,1) (1,2) (1,3) (2,1) (2,2) (2,3) (3,1) (3,2) (3,3)\}$$

5.4.3. Trial: An experiment can be repeated under essentially same conditions but may not yield similar results every time. Each repeat experiment is known as a trial. For example tossing of coin once, throwing of die once etc are single trials.

5.4.4. Event: The outcomes of these trials are known as events. For example, in case of tossing one coin the outcome is either Head (H) or Tail (T). So H and T are the events.

Some examples of events are

- Passengers arriving for check-in two hours before flight time.
- Getting a multiple of 3 in a single toss of a dice.
- Getting atleast two heads while tossing a coin twice.

An event may be

(a) Simple or elementary: if it contains only one outcome or one sample point.

(b) Composite or compound: if it contains more than one outcome

(c) Impossible: if it has no outcome

(d) Sure: If it consists of all possible outcomes and would definitely occur when the experiment is performed. The sample space is an example of a sure event.

5.4.1. Rules of Event Operations

- **Intersection of Events:** Intersection of two events say A and B indicates the common elements of A and B and expressed as

For example, consider the experiment of throwing a dice. The sample space S is

$$S = \{1, 2, 3, 4, 5, 6\}$$

Now, let A = Event that the number on the dice is even.

B = Event that the number on the dice is a multiple of 3.

Then A = {2, 4, 6}

B = {3, 6}

and $A \cap B = \{6\}$

Using a venn diagram, this can be represented as

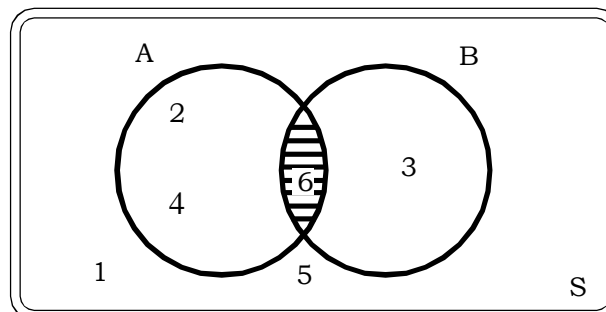


Figure 5.8

$$A \cap B = \{6\}$$

- Union of Events:** Union of two events $A \cup B$, indicates all the elements which belong either to A or to B given that they are mutually exclusive. Union of A and B is symbolized as $A \cup B$.

In the above experiment of tossing a dice, suppose we consider the same events A & B. Their union denoted by $A \cup B$ is given as

$$A \cup B = \{2, 3, 4, 6\}$$

A venn-diagram representation of $A \cup B$ is as follows:

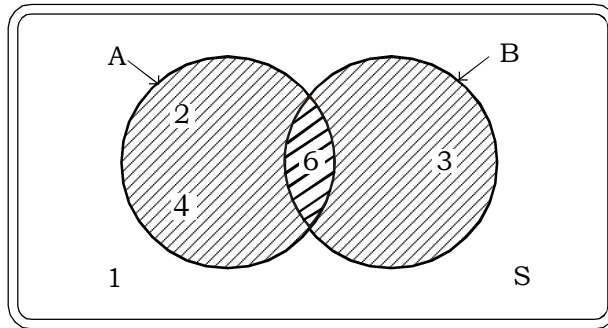


Figure 5.9

Union of two events

- Complement of an Event:** Complement of a particular event is the other event present in the same sample space. Complement of an event A is denoted by \bar{A} or A^c .

In the same example, since

$$A = \{2, 4, 6\}$$

$$A^c = \{1, 3, 5\}$$

and $\therefore B = \{3, 6\}$

$$B^c = \{1, 2, 4, 5\}$$

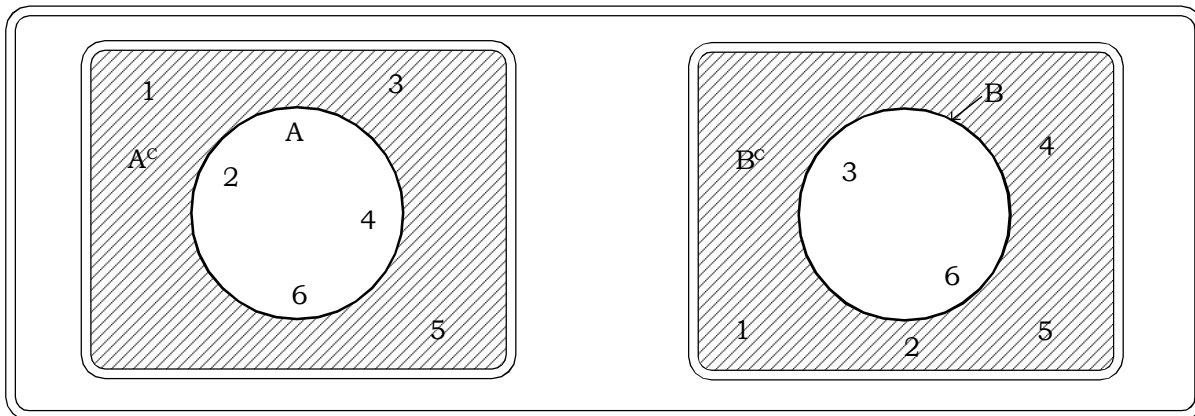


Figure 5.10

$$A^c = \{1, 3, 5\}$$

Figure 5.11

$$B^c = \{1, 2, 4, 5\}$$

5.4.4.2 Types of Events

Events can also be of the following different types:

- **Exhaustive Event:** Events will be called exhaustive event when they totally include all the possible outcomes of a random experiment. For example if a coin is tossed, Head (H) and Tail (T) are the two possible outcomes and hence exhaustive number of cases is 2.

Another example is if we draw a card from a pack of 52 cards, there would be 52 exhaustive cases or possible events.

- **Mutually Exclusive Event:** Mutually exclusive events are those events, which do not occur simultaneously. Occurrence of one indicates the absence of the other. Again it can be explained in terms of the outcome of a coin tossing. In tossing a coin either Head or Tail will occur. Both will not occur simultaneously, so the events Head and Tail are mutually exclusive. Mathematically, if two events A & B are mutually exclusive then

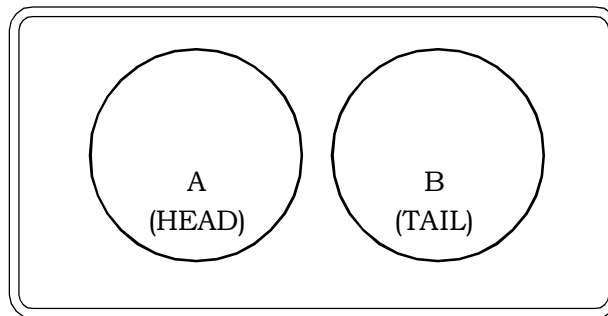


Figure 5.12

Two Mutually Exclusive Events

A Venn diagram depicting two mutually exclusive events A & B is shown in fig. 5.12.

- **Independent and Dependent Events:** Two or more events are said to be independent if the occurrence of one has no effect on the occurrence or non-occurrence of the others. For example in tossing a coin twice, the occurrence of head or tail in the first toss will be completely independent of the occurrence of head or tail in the second. As another example, if two children are born in a family, the birth of the first and the second are independent of each other. Events, which are not independent, are said to be dependent. For example, if a box contains 4 balls, 2 white and 2 red and two balls are drawn without replacement one after another. Then the event of getting a red ball in the second draw will depend on the event in the first draw.
- **Equally Likely Events:** If all the events have equal chance of occurrence, then the events are known as equally likely events. For example, again by bringing the case of coin tossing, the possibility of occurrence of head and tail are equal while tossing a coin.
- **Favorable Events:** The number of cases favorable to an event in a trial is the favorable cases of an event. For example if we throw a die, the possibility of getting sum 4 is

(1,3) (3,1) (2,2) i.e. 3

3 favorable cases for this particular event.

- **Complementary Event:** Two events are said to be complementary to each other if they are mutually exclusive and exhaustive. For example in case of the throwing of a die, occurrence of even number (2,4,6) and odd number (1,3,5) are complementary to each other.

- **Simple and Compound Event:** In case of a simple event, the probability of occurrence of only one event is considered, while in case of compound event more than one are considered. For example there are 5 red and 5 black pens in a box. If we want to know what is the probability of getting one black pen, it is a simple event. Again if we want to extend for the second trail and want to know the probability of one black and one red pen then it will be the case of a compound event.

5.5 VARIOUS DEFINITIONS OF PROBABILITIES

5.5.1 The Theoretical Definition of Probability

Let $e_1, e_2, e_3, \dots, e_n$ represent n outcomes of the sample space S

Thus $S = \{e_1, e_2, e_3, \dots, e_n\}$

Associated with each simple event $\{e_i\}$, $i = 1, 2, \dots, n$ we may assign a real number called the probability of $\{e_i\}$, denoted by $P(e_i)$. These probabilities satisfy the following conditions:

(a) $P(e_i)$ lies between 0 and 1, i.e.

$$0 \leq P(e_i) \leq 1$$

(b) $P(e_1) + P(e_2) + \dots + P(e_n) = 1$. The sum of all probabilities is unity.

(c) The probability of the impossible event is zero $P(\phi) = 0$

If A is an event, then probability of the event A denoted by $P(A)$ is the sum of the probabilities assigned to the simple events that comprise event A .

Thus if $A = \{a_1, a_2, a_3\}$

then $P(A) = P(a_1) + P(a_2) + P(a_3)$

There are different theories/definitions/approaches of probability. Broadly, the four common approaches of probability theory are:

- The Classical theory of Probability
- Statistical/Empirical/Relative Frequency definition of probability
- Axiomatic or Modern Probability Approach
- Subjective approach

All four approaches are now described.

5.5.2 The Classical Theory of Probability:

The classical theory of probability, also called 'a priori' probability, is the simplest definition of probability. In this approach, probability is defined as the ratio of the number of favorable cases of a certain event to the total number of cases possible.

Thus, probability of an event A , denoted by $P(A)$ is defined as:

$$P(A) = \frac{\text{Favourable number of cases of an event}}{\text{Total number of cases or number of exhaustive cases}}$$

Thus, if E – event of getting atleast one head in two tosses of a coin

The sample space = {HH, HT, TT, TH}

& E = {HH, HT, TH}

Thus by the classical definition of probability,

$$P(E) = \frac{\text{Favorable no. of cases}}{\text{Total no. of cases}} = \frac{3}{4}$$

Alternatively, assume there are 'N' elements, which can be subdivided into two groups' viz. favorable (n) cases and non-favorable (N – n) cases. Then the probability of favorable cases can be calculated as

$$P = \frac{n}{N}$$

and the probability of non-favorable cases is

$$1 - P = \frac{N - n}{N}$$

5.5.3 Relative Frequency or Empirical Approach of Probability:

The classical definition was found to have the following limitations.

- (i) When n, the total number of trials or exhaustive cases is infinite, the classical definition could not be applied.
- (ii) When all outcomes of the random experiment are not equally likely, this approach was not found suitable.

To overcome the shortcomings of the classical approach, the relative frequency approach was developed.

Relative frequency is estimated as the ratio of the number of occurrences of an event to the total number of times an experiment is repeated. The keyword in this approach is 'repeated'. For example we may say that there is 95% chance that our flight will be on time. This will be based on past experiences, which may be considered as repetitions of the experiment. Thus the ratio of past flights on time to the total no. of flights is believed to be approximately 95%. This is the idea behind the empirical or relative frequency approach to probability.

This approach approximates the probability of an event by calculating the proportion of time or the relative frequency with which the event has occurred over a finite number of repetitions of the experiment under identical conditions.

The formal definition of probability, according to the relative frequency approach, as given by R. Von Mises is as follows.

If an experiment is repeated n times, under essentially identical conditions, and if, out of these trials, an event A occurs m times, then the probability that A occurs is given by

$$P(A) = \lim_{n \rightarrow \infty} \frac{m}{n}, \text{ provided the limit exists}$$

This definition has the following shortcomings

- (i) When the number of trials become indefinitely large, the experimental conditions may no longer remain identical.
- (ii) The relative frequency may not attain a unique value no matter how large the total number of trials.
- (iii) This definition also does not facilitate any mathematical treatment of probability

This approach is used as an approximation when the number of trials is large. In fact the larger the number of trials, the better the approximation. This approach is also called empirical because the probability of an event is obtained by actual experimentation.

For Example: If 500 heads come from tossing of 1000 coins, then the probability of head is 0.5. If 600 heads come from the toss of 1500 the relative probability being = 0.4

5.5.4 Axiomatic Approach

Yet another approach of probability is the axiomatic approach. This approach is given by a Russian mathematician A. Kolmogorov in his book 'Foundations of Probability' in 1933. He defined probability as a function of the outcomes of an experiment, under certain restrictions known as postulates or axioms of probability. Through the axioms some rules are provided which help defining relationships between abstract entities. According to this approach the probability of an event A i.e. $P(A)$ satisfies the following axioms:

- (i) Axiom of Positiveness

$$0 \leq P(A) \leq 1$$

- (ii) Axiom of Certainty

If S denotes the sample space

$$P(S) = 1$$

- (iii) Axiom of Additivity

If A_1, A_2, \dots, A_n are n mutually exclusive events i.e. $A_i \cap A_j = \phi$, ϕ being a null set, then

$$i \neq j$$

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n) = \sum_{i=1}^n P(A_i)$$

The first axiom implies that the probability of any event is always a non-negative number less than or equal to unity.

The second axiom states that the probability of an event that is sure to occur must be equal to unity.

The third axiom gives a basic rule of addition of probabilities when the events are mutually exclusive.

The theoretical definition of probability is based on the above three axioms of the modern approach to probability.

5.5.5 Subjective Approach:

The third approach or a concept of probability is the subjective approach of probability. In this case depending on the available evidence probabilistic statements are made. Therefore it depends on the belief of a person who is giving the probability statement. For example an ex-athlete may be asked about the probability that a particular new athlete will achieve the first position in the race. From the past experience the ex-athletic can make a probability statement about the chances of the new athlete. This approach is thus based on an individual assessment or judgement about a random phenomenon, as the name suggests.

Example 5.8: Suppose a die is rolled. Find the probabilities of the following events:

- (i) The number is a multiple of 2
- (ii) The number is odd.

Solution:

First we the sample space when a die is rolled

$$S = \{1, 2, 3, 4, 5, 6\}$$

- (i) Let A: the number is a multiple of 2.

$$\text{Then } A = \{2, 4, 6\}$$

$$\text{Thus } P(A) = P(2) + P(4) + P(6) = \frac{3}{6} = \frac{1}{2}$$

- (ii) Let B: The number is odd

$$\text{Then } B = \{1, 3, 5\}$$

$$\text{Thus } P(B) = P(1) + P(3) + P(5)$$

$$= \frac{3}{6} = \frac{1}{2}$$

Example 5.9: A box contains 2 red, 5 white and 6 blue balls. What is the probability that out of two balls drawn, one is white and one is blue?

Solution:

The total number of balls = 2 + 5 + 6 = 13

2 balls can be drawn from it by ${}^{13}C_2$ ways. Out of 5 white balls one ball can be drawn in 5C_1 ways. Out of 6 blue balls one blue ball can be drawn 6C_1 ways.

Since the cases are related to each other, total no. of favorable cases is ${}^5C_1 \times {}^6C_1$.

$$\text{The required probability} = \frac{{}^5C_1 {}^6C_1}{{}^{13}C_2} = \frac{5}{13}$$

Example 5.10: A company is planning to make a committee of 3 people, which would consist of people from different department of the company itself. The number of officers nominated from production department is 4, from purchase department 5 and from sales 3. What is the probability that there must be one from each category?

Solution:

The total no. of nominated officers are = $4 + 5 + 3 = 12$

Out of 12 people 3 people can be selected by ${}^{12}C_3$ ways

Here the favorable cases are

$${}^4C_1 \times {}^5C_1 \times {}^3C_1$$

The required probability that the committee consists of one person from each category is:

$$\frac{{}^4C_1 \times {}^5C_1 \times {}^3C_1}{{}^{12}C_3} = \frac{4 \times 5 \times 3}{220} = \frac{3}{11}$$

Example 5.11: What is the probability that a leap year selected at random will contain 53 Sundays?

Solution:

A leap year consists of 366 days. This comprises of 52 weeks exact and 2 extra days. The two extra days could be any one of the following combinations:

- (i) (Sun, Mon)
- (ii) (Mon, Tues)
- (iii) (Tues, Wed)
- (iv) (Wed, Thurs)
- (v) (Thurs, Fri)
- (vi) (Fri, Sat)
- (vii) (Sat, Sun)

The possibility of a Sunday is in two instances: (Sun, Mon) and (Sat, Sun)

The required probability = $\frac{2}{7}$

Example 5.12: The probabilities that a student would receive an A, B, C, D, F or incomplete in a test are 0.05, 0.15, 0.20, 0.25, 0.30 & x.

- (i) Find x
- (ii) Find the probability that he will get a A.
- (ii) Find that probability that he will get almost C.
- (iii) Find the probability that he will get at least a C.

Solution:

- (a) Given $P(A) = 0.05$
- $P(B) = 0.15$
- $P(C) = 0.20$
- $P(D) = 0.25$
- $P(F) = 0.30$
- $P(\text{incomplete}) = x$

Thus $P(A) + P(B) + P(C) + P(D) + P(F) + P(\text{incomplete (I)}) = 1$

Thus, $P(I) = 1 - (0.05 + 0.15 + 0.20 + 0.25 + 0.30)$

$$\begin{aligned}\Rightarrow x &= 1 - 0.95 \\ &= 0.05\end{aligned}$$

(ii) $P(A) = 0.05$

$$\begin{aligned}\text{(iii) } P(\text{Getting at most a C}) &= P(I) + P(F) + P(D) + P(C) \\ &= 0.05 + 0.30 + 0.25 + 0.20 \\ &= 0.8\end{aligned}$$

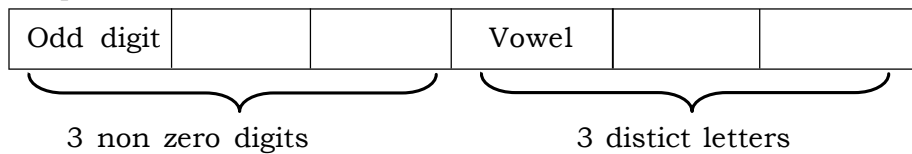
(iv) Probability of getting at least a C

$$\begin{aligned}P(\text{getting at least a C}) &= P(A) + P(B) + P(C) \\ &= 0.05 + 0.15 + 0.20 \\ &= 0.4\end{aligned}$$

Example 5.13: If each licence plate contains 3 distinct non zero digits followed by 3 distinct letters, find the probability that, if a license plate is picked at random, the first digit will be odd and the first letter will be a vowel.

Solution:

There are six positions in all



The 9 non zero digits are 1, 2, 3, 4, 5, 6, 7, 8 and 9 out of which 5 odd digits are viz 1, 3, 5, 7, and 9.

There are 26 alphabets in all & 5 vowels.

The total number of ways in which 3 distinct non zero digits and 3 distinct letters can be arranged is ${}^9P_3 \cdot {}^{26}P_3$.

The number of ways in which the first digits can be odd is 5P_1 ways

And since these are 5 vowels, the no of ways in which the first letter will be a vowel is also 5P_1 ways.

In the remaining two places for digits, two digits can be arranged from the 8 digits in 8P_2 ways.

And in the remaining two places for the letter; the numbers of ways of arranging two letters from the 25 letters is ${}^{25}P_2$

$$\begin{aligned}\text{Thus the required probability} &= \frac{{}^5P_1 \cdot {}^5P_1 \cdot {}^8P_2 \cdot {}^{25}P_2}{{}^9P_3 \cdot {}^{26}P_3} \\ &= \frac{25}{234}\end{aligned}$$

Example 5.14: There are 8 people at a picnic and their ages are as follows: 15, 5, 2, 20, 7, 30, 40, 23. If three people are picked at random, what is the probability that their combined ages will exceed 27 years?

Solution:

Their combined ages would not exceed 27 years in the following cases:

$$15 + 5 + 2, 15 + 5 + 7, 15 + 7 + 2, 20 + 5 + 2, 5 + 2 + 7.$$

i.e. 5 cases in all.

The total no of combinations of 3 possible with 8 people are

$${}^8C_3 = 56$$

Out of these 56 cases, in 5 cases the combined ages would not exceed 27 years.

∴ Number of cases in which the combined ways would exceed 27 years is $56 - 5 = 51$

∴ Required probability that the combined ages would exceed 27 years = $\frac{51}{56}$

Example 5.15: A box contains 8 good pens, 4 pens with minor defects and 3 pens with major defects. Four pens are picked at random without replacement. Find the probability that

- (i) All pens are defective
- (ii) Atleast one pen is good
- (iii) Exactly 2 pens are good
- (iv) One pen is good, one has a minor defect and 2 have a major defect
- (v) One pen is good and 3 have major defects.

Solution:

The total number of ways in which 4 pens can be drawn at random from 15 pens is

$${}^{15}C_4 = \frac{15!}{4!11!} = \frac{15.14.13.12}{4.3.2.1} = 1365 \text{ ways}$$

- (i) 4 defective pens can be selected out of 7 defective pens in

$${}^7C_4 = \frac{7.6.5.4}{3.2.1} = 35 \text{ ways}$$

Thus, probability that all 4 pens are defective

$$= \frac{35}{1365} = \frac{1}{39}$$

- (ii) P (atleast one pen is gold)

$$= 1 - P(\text{all the pens are defective})$$

$$= 1 - \frac{1}{39} = \frac{38}{39}$$

(iii) Exactly two good and two bad pens may be selected in

$${}^8C_2 {}^7C_2 = \frac{8 \cdot 7}{2} \frac{7 \cdot 6}{2} = 588 \text{ ways}$$

∴ Required probability that exactly 2 pens are good

$$= \frac{588}{1365} = \frac{196}{455}$$

(iv) One good pen one pen with a minor defect and 2 pens with a major defect can be chosen in

$${}^8C_1 {}^4C_1 {}^3C_2 = 8 \times 4 \times 3 = 96 \text{ ways.}$$

$$\text{Thus, the required probability} = \frac{96}{1365} = \frac{32}{455}$$

(v) One good pen, and three pens with major defect can be selected in

$${}^8C_1 {}^3C_3 = 8 \text{ ways}$$

$$\text{The required probability} = \frac{8}{1365}$$

Example 5.16: A six digit number is formed with the digits 1, 2, 3, 4, 5, 7 with no repetitions. What is the probability that

(i) The number is even?

(ii) The number is divisible by 5, that is, the units' digit is 5?

Solution:

(i) For the number to be even last two digits must be either 2 or 4 out of a total of 6 digits.

$$\therefore \text{Required probability} = \frac{2}{6} = \frac{1}{3}$$

(ii) The number would be divisible by 5, if the last digit is 5 i.e. 1 case out of a total of 6 cases

$$\therefore \text{The required probability} = \frac{1}{6}$$

Example 5.17: Twenty people have been invited to a party. Find the number of handshakes that would take place if each person shakes hands with everyone else in the group.

Solution: Since there are twenty guests and each guest has to shake hands with each other; we have to find combinations of 2 out of 20. Thus, the required number of handshakes.

$$\begin{aligned} &= {}^{20}C_2 \\ &= \frac{20 \cdot 19 \cdot 18!}{2! \cdot 18!} \\ &= 190 \text{ ways} \end{aligned}$$

Example 5.18: Sixteen college graduates have applied for 5 vacancies in an organization. In how many ways can the organization make the five offers to the sixteen graduates?

Solution: Number of ways in which the organization can make five offers to the sixteen graduates

$$= {}^{16}C_5 = \frac{16!}{5!.11!} = \frac{524160}{120} = 4368 \text{ ways}$$

Example 5.19: Five couples occupy 10 seats in a row at random. What is the probability that all the ladies are sitting next to each other?

Solution:

10 people can arrange themselves in 10 seats in $10!$ ways.

Considering the two ladies as one, 6 people can sit in a row in $6!$ Ways.

Further, the ladies can rearrange themselves in $5!$ ways

Therefore, the required probability

$$= \frac{6!5!}{10!}$$

$$= \frac{6!.5.4.3.2}{10.9.8.7.6!} = 0.02$$

5.6 LAWS & THEOREMS OF PROBABILITY

5.6.1 Additive Law:

If A and B are two events, then $P(A \cup B)$ which indicates the occurrence of either event A or event B or both is equal to the sum of their individual probabilities minus the probability of their simultaneous occurrence, if they are not mutually exclusive i.e.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If A and B are mutually exclusive then $P(A \cap B) = 0$. Thus

$$P(A \cup B) = P(A) + P(B)$$

Remarks

- (i) The event $A \cup B$ denotes the occurrence of either A or B or both, thus implying the occurrence of at least one of the two events.
- (ii) $A \cap B$ is a compound event denoting simultaneous occurrence of A & B
- (iii) Also,

P (none of the events A and B occur simultaneously)

$$\text{i.e. } P(\bar{A} \cap \bar{B}) = 1 - P(A \cup B)$$

Generalization

The additive law can be generalized to more than two events.

In case of three events A, B and C in a sample space S, the probability of occurrence of at least one of them is given by

$$\begin{aligned} P(A \cup B \cup C) &= P[A \cup (B \cup C)] \\ &= P(A) + P(B \cup C) - P[A \cap (B \cup C)] \\ &= P(A) + P(B) + P(C) - P(B \cap C) - P[(A \cap B) \cup (A \cap C)] \\ &= P(A) + P(B) + P(C) - P(B \cap C) - P(A \cap B) - P(A \cap C) + P(A \cap B \cap C) \end{aligned}$$

Thus

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(B \cap C) - P(A \cap B) - P(A \cap C) + P(A \cap B \cap C)$$

If A, B, and C are mutually exclusive

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

Also, the probability that none of the events A, B and C occur simultaneously.

$$P(\bar{A} \cap \bar{B} \cap \bar{C}) = 1 - P(A \cup B \cup C)$$

Thus, for n events A_1, A_2, \dots, A_n

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = \sum P(A_i) - \sum_{i \neq j} P(A_i \cap A_j) + \sum_{i \neq j \neq k} P(A_i \cap A_j \cap A_k) + (-1)^n P(A_1 \cap A_2 \cap \dots \cap A_n)$$

$$\text{and } P(\bar{A}_1 \cap \bar{A}_2 \cap \dots \cap \bar{A}_n) = 1 - P(A_1 \cup A_2 \cup \dots \cup A_n)$$

And if A_1, A_2, \dots, A_n are mutually exclusive

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = \sum_{i=1}^n P(A_i)$$

Remark:

The event $A \cup B$ is also denoted by $A + B$ and the event $A \cap B$ is also denoted by AB .

5.6.2 Conditional Probability and Multiplication Law of Probability

Conditional probability is the probability of an event occurring under the consideration that another event has already taken place. Alternatively, it is the probability of an event, given some condition.

For example, suppose we want to know the probability of event 'A' given that another 'B' has already occurred. Symbolically it can be presented as $P(A/B)$ and read as "probability of A given B."

Mathematically, the conditional probability of an event A given the occurrence of an event B is defined by

$$P(A/B) = \frac{P(AB)}{P(B)}, \text{ Provided } P(B) \neq 0$$

$P(A/B)$ is not defined if $P(B) = 0$

This formula can be rewritten as

$$P(A \cap B) = P(B) \cdot P(A/B)$$

This is called the general multiplication law of probability, and gives the probability that the two events A and B would occur simultaneously.

For three events A, B and C

$$P(A \cap B \cap C) = P(A \ B \ C) = P(A) P(B/A) P(C/AB)$$

In general, for n events A_1, A_2, \dots, A_n .

$$P(A_1 \cap A_2 \dots \cap A_n) = P(A_1 \dots A_n) = P(A_1) P(A_2/A_1) P(A_3/(A_1A_2)) \dots P(A_n/A_1 \dots A_{n-1})$$

5.6.3 Theory of Independence

Two events say A and B are called independent if

- $P(A/B) = P(A)$ i.e. B does not effect the occurrence of A
- $P(B/A) = P(B)$ i.e. A does not effect the occurrence of B

Therefore from the result, $P(AB) = P(A/B) P(B)$,

if A & B are independent we have

$$P(A \cap B) = P(A) P(B)$$

Thus, two events A and B are said to be independent if the probability of simultaneous occurrence of A and B is equal to the product of their individual probabilities i.e.

$$P(A \ B) = P(A) \cdot P(B)$$

Generalization

In general if A_1, A_2, \dots, A_n are independent events the multiplicative law of probability becomes

$$P(A_1 \cap A_2 \dots \cap A_n) = P(A_1) P(A_2) \dots P(A_n)$$

5.6.4 Pair-wise and Mutual Independence

Three events A, B, and C are said to be mutually independent if the following conditions are satisfied simultaneously

- (i) $P(AB) = P(A) P(B)$
- (ii) $P(BC) = P(B) P(C)$
- (iv) $P(AC) = P(A) P(C)$
- (v) $P(ABC) = P(A)P(B)P(C)$

When only the first three conditions are satisfied and the fourth condition is not satisfied, then the events are said to be pair-wise independent.

Thus, mutually independent events are always pair-wise independent, but pair-wise independent events may or may not be mutually independent.

5.6.5 The Theorem of Total Probability and The Baye's Theorem

Baye's theorem is a simple mathematical formula used for calculating special type of conditional probabilities. The theorem is associated with the name of the famous English statistician Thomas Bayes (1702 – 1761). The probability of an event A conditional on another event B is different from the probability of B conditional to A. However, a definite relation is there between these two types of conditional probability. Baye's theorem deals with this basic relationship.

A practical example: Suppose there are four machines each producing a certain percentage of defective items. If we pick up an item at random from a machine, we are often interested in the probability that the item is defective.

Bayes theorem treats the reverse problem. It aims to find the probability that the item came from a particular machine, given that the item is defective.

The Theorem of Total Probability

This is given by as follows:

Suppose B_1, B_2, \dots, B_n are mutually exclusive and exhaustive events and A is any event, all the events having a positive probability.

Then

$$P(A) = P(A/B_1) P(B_1) + P(A/B_2) P(B_2) + \dots + P(A/B_n) P(B_n)$$

A generalization of this theorem gives the Bayes Theorem.

The Bayes Theorem

$$P(B_i/A) = \frac{P(A/B_i) P(B_i)}{P(A/B_1) P(B_1) + \dots + P(A/B_n) P(B_n)}, \quad i = 1, 2, \dots, n$$

The probabilities $P(B_1), \dots, P(B_n)$ are called a priori or prior probabilities and the probabilities $P(B_1/A), P(B_2/A), \dots, P(B_n/A)$ are called a-posterior or simply posterior Probabilities.

The following examples now look at the applications of the laws & theorems discussed in this section.

Example 5.20: In a college survey of 100 MBA students, the combination of subjects they have taken during Autumn Semester are as follows:

Subjects Opted by MBA students in a college

Human Resource	80
Operation Research	75
Production	60
Human Resource & Operation Research	40
Human Resource & Production	50
Operation Research & Production	30

Find out how many students have taken all the three subjects.

Solution:

Let the events be symbolized as follows: A student opting for

Subjects	Symbol	No. of Students
Human Resource	A	80
Operation Research	B	75
Production	C	60
Human Resource & Operation Research	AB	40
Human Resource & Production	BC	50
Operation Research & Production	ABC	30

Then, the number of students who have taken all three subjects are

$$(A \cup B \cup C) \text{ or } A + B + C = [A] + [B] + [C] - [AB] - [AC] - [BC] + [ABC]$$

$$\Rightarrow 100 = 80 + 75 + 60 - 40 - 50 - 30 + [ABC]$$

$$\Rightarrow ABC = 5$$

There are 5 students who have taken all the three subjects.

Example 5.21: A salesman has a 60 percent chance of making a sale to each customer. The behavior of successive customers is independent. If two customers A and B enter, what is the probability that the salesman will make a sale to A or B? (MBA, DU, 1998)

Solution: By, the addition law, probability that the salesman will make a sale to A or B is

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) - P(AB) \\ &= 0.60 + 0.60 - (0.60)(0.60) \\ &= 1.2 - 0.36 = 0.84 \end{aligned}$$

Thus there is 84% chance that the salesman will make a sale to A or B.

Example 5.22: A newspaper seller is interested in finding out the chances of selling more than 90 copies of the newspaper. He observed from his past year record that out of 365 days, on 100 days he had sold 70 copies, 150 days 95 copies and 115 days 100 copies. What is the required probability?

Solution:

Sales of Newspaper

Sales	No. of Days
70	100
95	150
100	115
	365

The probabilities of selling more than 90 newspaper is given by $\frac{150+115}{365} = \frac{265}{365} = 0.73$

Example 5.23: The probability that India wins a cricket test match against Pakistan is, given to be $1/3$. If India and Pakistan play six test matches, what is the probability that:

- (i) India will lose all the six test matches?
- (ii) India will win at least one test match? (MBA, DU, 2000)

Solution:

Let

p - Probability that India wins a cricket test match = $\frac{1}{3}$

q - Probability that India losses a cricket test match = $\frac{2}{3}$

(i) Prob (India will lose all the 6 test matches) = $\left(\frac{2}{3}\right)^6 = 0.088$

(ii) P (India will win atleast one test match) = $1 - P$ (India will lose all the 6 test matches)
 = $1 - 0.088$
 = 0.912

Example 5.24: A company has two plants to manufacture scooters. Plant I manufacturer 80% of the scooters and plant II manufacturers 20%. At plant I, 85% scooters are rated as standard quality. At plant II, only 65% scooters are rated as standard quality.

- (i) What is the probability, that a customer obtains a standard quality scooter if he buys a scooter from the company?
- (ii) What is the probability that the scooter came from plant1, if it is known that the scooter is of standard quality? (MBA, DU, 1998)

Solution: Let

P_1 - The event that the Scooter is manufactured in Plant 1

P_2 - The event that the Scooter is manufactured in Plant 2

Q - The event that the Scooter is of standard quality

$$P(P_1) = 0.80$$

$$P(P_2) = 0.20$$

$$P(Q/P_1) = 0.85$$

$$P(Q/P_2) = 0.65$$

- (i) By applying the theorem of Total probability,

$$\begin{aligned} P(Q) &= P(P_1) P(Q/P_1) + P(P_2) P(Q/P_2) \\ &= (0.80)(0.85) + (0.20)(0.65) \\ &= 0.68 + 0.13 \\ &= 0.81 \end{aligned}$$

The probability that a customer obtains a standard quality scooter = 0.81

(ii) This example is a typical application of the Bayes Theorem.

$$\begin{aligned} P(P_1/Q) &= \frac{P(P_1)P(Q/P_1)}{P(P_1)P(Q/P_1) + P(P_2)P(Q/P_2)} \\ &= \frac{(0.80)(0.85)}{(0.80)(0.85) + (0.20)(0.65)} = \frac{0.68}{0.81} = 0.84 \end{aligned}$$

Thus there is 84% chance that standard quality scooter came from Plant 1.

Example 5.25: An automobile company has two emergency on road services- service A and service B. In case of calls, the probability that service A responds is 0.67 and the probability that service B respond is 0.85. The probability that either A or B responds is 0.98. Find the probability that both services respond to a call.

Solution:

$$P(A) = 0.67 = P(\text{A responds to a call})$$

$$P(B) = 0.85 = P(\text{B respond to a call})$$

$$P(A \cup B) = 0.98$$

We have to find $P(A \cap B)$, i.e. the probability that both services respond to a call.

By additive law of probability.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$\begin{aligned} P(A \cap B) &= P(A) + P(B) - P(A \cup B) \\ &= 0.67 + 0.85 - 0.98 \\ &= 0.54 \end{aligned}$$

Thus, there is a 54% chance that both services would respond to a call.

Example 5.26: The probability that a person goes for a movie on Saturday is $\frac{2}{3}$ and the probability that he goes for a play on Sunday is $\frac{4}{5}$. Probability that he goes either to a movie or to a play is $\frac{8}{9}$. Find:

- (i) the probability that he attends both
- (ii) the probability that he attends none

Solution:

$$P(\text{Person attending a movie}) = \frac{2}{3}$$

$$P(\text{Person attending a play}) = \frac{4}{5}$$

$$\begin{aligned} \text{(i) } P(\text{Person attends either a movie or a play}) &= \frac{2}{3} + \frac{4}{5} - \frac{8}{9} \\ &= 0.66 + 0.8 - 0.89 \\ &= 0.57 \end{aligned}$$

Thus there is 57% chance that the person will end up going for both movie as well as a play.

$$\begin{aligned}
 \text{(ii) } P(\text{Person attends neither movie nor play}) & \\
 &= 1 - P(\text{Person attends both movie \& play}) \\
 &= 1 - 0.57 \\
 &= 0.43
 \end{aligned}$$

Thus, there is 43% chance that the person will go neither for a movie or a play.

Example 5.27: In a cafe, the probability that a customer buys a coffee is 0.43 and the probability that he/she buys a tea is 0.2. A customer does not buy both. What is the probability that a customer buys either tea or coffee?

Solution: Given:

$$\begin{aligned}
 P(\text{a customer buys tea}) &= 0.2 \\
 P(\text{a customer buys coffee}) &= 0.43 \\
 P(\text{a customer buys both tea and coffee}) &= 0 \\
 P(\text{A customer buys either tea or coffee}) &= P(\text{Tea}) + P(\text{Coffee}) \\
 &= 0.3 + 0.2 \\
 &= 0.5
 \end{aligned}$$

Example 5.28: In an automobile showroom, the sales executive from past experience has found out that the probability that a customer buys a small utility vehicle is 0.3, the probability that a customer buys a mid sized car is 0.15 and the probability that a customer buys a multi utility vehicle is 0.09. What is the probability that the customer buys any one of the above three types of vehicle?

Solution: This is an application for the additive law for three events that are mutually exclusive.

$$\begin{aligned}
 P(\text{a customer buys small utility vehicle}) &= 0.3 \\
 P(\text{a customer buys mid sized}) &= 0.15, P(\text{multi-utility vehicle}) = 0.09 \\
 P(\text{a customer buys either one of the three}) &= 0.3 + 0.15 + 0.09 \\
 &= 0.54
 \end{aligned}$$

Example 5.29: The probability that a light bulb will last beyond 200 hours is 0.7 and the probability that it will last beyond 250 hours is 0.28. Given that the bulb last beyond 200 hours, what is the probability that it lasts beyond 250 hours?

Solution:

Let B_{200} = event that the bulb lasts beyond 200 hours.

B_{250} = event that the bulb lasts beyond 250 hours.

We need to find $P(B_{250} / B_{200})$

By definition of conditional probability

$$\begin{aligned}
 P(B_{250} / B_{200}) &= \frac{P(B_{250}B_{200})}{P(B_{200})} \\
 &= \frac{P(B_{250})}{P(B_{200})} \\
 &= \frac{0.28}{0.7} = 0.4
 \end{aligned}$$

Thus, the required probability that the bulb lasts beyond 250 hours = 0.4.

Example 5.30: A door-to-door salesman of a direct selling product makes house calls to different households to sell the product. The probability that the salesman finds a person at home is 0.75. Given that he finds a person at home, the probability that he's able to make a sale is 0.28. Find the probability that the salesman finds a person at home and also manages to make a sale.

Solution:

$$P(\text{person is at home}) = 0.75$$

$$P(\text{person buys the product/he is at home}) = 0.28$$

$$\begin{aligned} P(\text{person is at home \& person buys the product}) \\ &= P(\text{person is at home}) P(\text{person buys the product/he is at home}) \\ &= (0.75)(0.28) = 0.21 \end{aligned}$$

Example 5.31: The probability that a book will be favorably reviewed by two critics is $\frac{5}{7}$ and $\frac{4}{7}$ respectively. If the reviews are independent, find the probability that either of them will give a favorable review.

Solution:

Let C_1 : event that the book is favorably reviewed by first critic

C_2 : event that the book is favorably reviewed by second critic

$$\text{Let } P(C_1) = \frac{5}{7}$$

$$P(C_2) = \frac{4}{7}$$

The probability that either of them would give a favourable review:

$$P(C_1 \cup C_2) = P(C_1) + P(C_2) - P(C_1)P(C_2); C_1 \text{ and } C_2 \text{ are independent}$$

$$\begin{aligned} &= \frac{5}{7} + \frac{4}{7} - \frac{20}{49} \\ &= 0.71 + 0.57 - 0.41 \\ &= 0.87 \end{aligned}$$

Example 5.32: A problem in statistics is given to three students whose chances of solving it are $\frac{1}{2}$, $\frac{1}{3}$ and $\frac{1}{4}$. What is the probability that the problem will be solved?

Solution:

Let

$$S_1 = 1^{\text{st}} \text{ student solves the problem \& } P(S_1) = \frac{1}{2} = 0.5$$

$$S_2 = 2^{\text{nd}} \text{ student solves the problem \& } P(S_2) = \frac{1}{3} = 0.33$$

$$S_3 = 3^{\text{rd}} \text{ student solves the problem \& } P(S_3) = \frac{1}{4} = 0.25$$

The probability that the problem is solved:

$$P(S_1 \cup S_2 \cup S_3) = P(S_1) + P(S_2) + P(S_3) - P(S_1S_2) - P(S_2S_3) - P(S_1S_3) + P(S_1S_2S_3) \text{ (using additive law for three events)}$$

Since the students try independently,

$$\begin{aligned} &= (0.5) + (0.33) + (0.25) - (0.5)(0.33) - (0.33)(0.25) - (0.25)(0.5) + (0.5)(0.25)(0.33) \\ &= 1.08 - 0.165 - 0.0825 - 0.125 + 0.04125 \\ &= 0.748 \end{aligned}$$

Thus, the probability that the problem is solved is 0.748.

Example 5.33: The probability that the stock market goes up on Tuesday is 0.64. Given that it goes up on Tuesday, the probability that it goes up on Wednesday is 0.35. What is the probability that it goes up on both days?

Solution:

Let T = event that the market goes up on Tuesday.

W = event that the market goes up on Wednesday.

Given

$$P(T) = 0.64$$

$$P(W/T) = 0.35$$

We have to find the probability that it goes up on both days i.e.

$$\begin{aligned} P(WT) &= P(W/T) P(T) \\ &= (0.35)(0.64) \\ &= 0.224 \end{aligned}$$

Example 5.34: An oil company has decided to drill two wells while exploring for oil in a certain area. The probability of striking oil in the first well is 0.25. Given that it struck oil in first well, the probability of striking oil in the second well is 0.82. Find the probability of striking oil in both the wells.

Solution:

Let W_1 – event of striking oil in the first well.

W_2 – event of striking oil in the second well

Given

$$P(W_1) = 0.25$$

$$P(W_2/W_1) = 0.82$$

We have to find the probability of striking oil in both the wells i.e.

$$\begin{aligned} P(W_1W_2) &= P(W_2/W_1) P(W_1) \\ &= (0.25)(0.82) \\ &= 0.205 \end{aligned}$$

Example 5.35: The probability that there will be fog in the morning is 0.3. Given that there is fog, the probability that a plane will take off on time is 0.43. Given that there is no fog, the probability that the plane will take off on time is 0.88. Find the probability that

- (i) There will be fog and the plane will take off on time.
- (ii) There will be no fog and the plane will take off on time
- (iii) The plane will take off on time

Solution:

Let F – event that there is fog in the morning.

\bar{F} – event that there is no fog in the morning

P – event that plane takes off on time.

\bar{P} – event that plane does not take off on time

Given

$$P(F) = 0.3$$

$$P(\bar{F}) = 0.7$$

$$P(P/F) = 0.43$$

$$P(P/\bar{F}) = 0.88$$

$$(i) P(FP) = P(P/F) P(F) = (0.43)(0.3) = 0.129$$

$$(ii) P(\bar{F}P) = P(P/\bar{F}) P(\bar{F}) = (0.88)(0.7) = 0.616$$

$$(iii) P(P) = P(FP) + P(\bar{F}P) = 0.129 + 0.616 = 0.745$$

Example 5.36: On a day in June, the probability that it will rain in Delhi is 0.3 and the probability that it will rain in Guwahati is 0.75. Assuming independence, find the probability that it will rain

- (i) In both Delhi & Guwahati
- (ii) Neither cities

Solution:

$$P(\text{rain in Delhi}) = 0.3$$

$$P(\text{rain in Guwahati}) = 0.75$$

$$(i) P(\text{rain in both Delhi \& Guwahati}) = P(\text{rain in Delhi}) P(\text{rain in Guwahati}) \\ = (0.3)(0.75) = 0.225$$

$$(ii) P(\text{rain in neither cities}) = 1 - P(\text{either in Delhi or Guwahati}) \\ = 1 - [P(\text{Delhi}) + P(\text{Guwahati}) - 0.225] \\ = 1 - [0.3 + 0.75 - 0.225] \\ = 1 - 0.825 = 0.175$$

Example 5.37: In a multiple choice test, a student either knows the answer or guesses both with equal probability. The probability that the student guesses the answer and it is correct is $\frac{1}{4}$. Find the probability that a student knows the answer, given that the answer is correct.

Solution:

Let A – the event that his answer is correct.

B_1 – the event that he guesses the answer

B_2 – the event that he knows the answer

$$P(B_1) = \frac{1}{2} = P(B_2)$$

$$P(A/B_1) = \frac{1}{4}$$

$P(A/B_2) = 1$ i.e. probability that the answer is correct when he knows the answer.

Applying The Baye's Theorem, the probability that the student knows the answer given that the answer is correct.

$$\begin{aligned} P(B_2/A) &= \frac{P(B_2)P(A/B_2)}{P(B_1)P(A/B_1) + P(B_2)P(A/B_2)} \\ &= \frac{\frac{1}{2} \cdot 1}{\frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot 1} = \frac{4}{5} = 0.8 \end{aligned}$$

Example 5.38: Mr. Ram speaks the truth in 3 out of 4 times, while Mr. Shyam speaks the truth in 4 out of 5 times. Find the probability that they will contradict each other in stating the fact. (MBA, DU, 1998).

Solution:

Let R – event that Mr. Ram speaks the truth

\bar{R} – event that Mr. Ram does not speaks the truth

Let S - event that Mr. Shyam speaks the truth

\bar{S} - event that Mr. Ram does not speaks the truth

$$P(R) = \frac{3}{4}, P(S) = \frac{4}{5}$$

$$P(\bar{R}) = \frac{1}{4}, P(\bar{S}) = \frac{1}{5}$$

The two of them would contradict each other when either of them do not speak the truth.

P (The two of them contradict each other in stating a fact)

$$\begin{aligned} &= P(R \cap \bar{S}) + P(\bar{R} \cap S) \\ &= \frac{3}{4} \times \frac{1}{5} + \frac{4}{5} \times \frac{1}{4} = 0.15 + 0.2 \\ &= 0.35 \end{aligned}$$

Example 5. 39: In a bolt factory machines A, B and C manufactures respectively 25%, 35% and 40%. Of the total of their output, 5,4 and 2 percent are defective bolts. A bolt is drawn at random from the product and is found to be defective. What are the probabilities that it was manufactured by machines A, B and C respectively?

Solution:

Let D – bolt is defective

A – bolt is manufactured by machine A

B – bolt is manufactured by machine B

C – bolt is manufactured by machine C

$$P(A) = 0.25$$

$$P(B) = 0.35$$

$$P(C) = 0.40$$

$$P(D/A) = 0.05$$

$$P(D/B) = 0.04$$

$$P(D/C) = 0.02$$

We have to find $P(A/D)$, $P(B/D)$ and $P(C/D)$

By applying Baye's Theorem

$$\begin{aligned} P(A/D) &= \frac{P(A)P(D/A)}{P(A)P(D/A) + P(B)P(D/B) + P(C)P(D/C)} \\ &= \frac{(0.25)(0.05)}{(0.25)(0.05) + (0.35)(0.04) + (0.4)(0.02)} \\ &= \frac{0.0125}{0.0125 + 0.014 + 0.008} = \frac{0.0125}{0.0345} \\ &= 0.36 \end{aligned}$$

$$\begin{aligned} P(B/D) &= \frac{P(B)P(D/B)}{P(A)P(D/A) + P(B)P(D/B) + P(C)P(D/C)} \\ &= \frac{.014}{0.0345} = 0.40 \end{aligned}$$

And similarly,

$$\begin{aligned} P(C/D) &= \frac{P(C)P(D/C)}{P(A)P(D/A) + P(B)P(D/B) + P(C)P(D/C)} \\ &= \frac{0.014}{0.0345} = 0.23 \end{aligned}$$

Thus, probability that it was manufactured by machine A = 0.36

Thus, probability that it was manufactured by machine B = 0.40

Thus, probability that it was manufactured by machine C = 0.23

Example 5.40: Three persons A, B and C are being considered for the appointment as Vice – Chancellor of a University whose chances of being selected for the post are in the proportion 4:2:3 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?

Solution:

Let D – democratization is introduced in the University.

A – A is selected as Vice Chancellor

B – B is selected as Vice Chancellor

C – C is selected as Vice Chancellor

$$P(A) = \frac{4}{9}, P(B) = \frac{2}{9}, P(C) = \frac{3}{9}$$

$$P(D/A) = \frac{3}{10}, P(D/B) = \frac{5}{10}, P(D/C) = \frac{8}{10}$$

Thus, by theorem of total probability, the required probability is

$$\begin{aligned} P(D) &= P(DA \cup DB \cup DC) \\ &= P(DA) + P(DB) + P(DC) \\ &= P(A) P(D/A) + P(B) P(D/B) + P(C) P(D/C) \\ &= \frac{4}{9} \times \frac{3}{10} + \frac{2}{9} \times \frac{5}{10} + \frac{3}{9} \times \frac{8}{10} = \frac{23}{45} \\ &= 0.511 \end{aligned}$$

Example 5.41: 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

- (i) Both of them will be selected.
- (ii) Only one of them will be selected.
- (iii) None of them will be selected.

(MBA, DU, 1999, M. Com, Madurai Kamaraj, Nov., 2003)

Solution:

Let H – husband is selected for the post, \bar{H} = husband is not selected for the post

W – Wife is selected for the post and \bar{W} = wife is not selected for the post

$$P(H) = \frac{1}{7}, P(W) = \frac{1}{5}. \text{ And } P(\bar{H}) = \frac{6}{7}, P(\bar{W}) = \frac{4}{5}$$

$$(i) P(HW) = P(H) P(W)$$

$$= \frac{1}{7} \times \frac{1}{5} = \frac{1}{35}$$

$$(ii) P(H\bar{W}) + P(\bar{H}W) = P(\text{only one of them will be selected})$$

$$= P(H) P(\bar{W}) + P(\bar{H}) P(W)$$

$$= \frac{1}{7} \times \frac{4}{5} + \frac{6}{7} \times \frac{1}{5} = \frac{10}{35}$$

(iii) P (none of them will be selected)

$$\begin{aligned} &= P(\overline{H}\overline{W}) = P(\overline{H})P(\overline{W}) \\ &= \frac{6}{7} \times \frac{4}{5} = \frac{24}{35} \end{aligned}$$

Example 5.42: The odds against student X solving a Business Statistics problem are 8:6 and odds in favor of student Y solving the same problem is 14:16.

- (i) What is the chance that the problem will be solved if both try?
 (ii) What is the probability that they both, working independently of each other, solve the problem?
 (iii) What is the probability that neither solves the problem? **(MBA, M.K. Univ., Nov., 2003)**

Solution:

Let X – X solves the business statistics problem

Y – Y solves the business statistics problem

$$P(\overline{X}) = \frac{8}{14}, \quad P(X) = \frac{6}{14}$$

$$P(Y) = \frac{14}{30}, \quad P(\overline{Y}) = \frac{16}{30}$$

(i) P (problem will be solved if they both try)

$$\begin{aligned} &= P(X \cup Y) \\ &= P(X) + P(Y) - P(X)P(Y) \\ &= \frac{6}{14} + \frac{14}{30} - \left(\frac{6}{14}\right)\left(\frac{14}{30}\right) \\ &= (0.43) + (0.47) - 0.2 \\ &= 0.7 \end{aligned}$$

(ii) P (both solve the problem independently)

$$\begin{aligned} &= P(X \cap Y) \\ &= P(X)P(Y) \\ &= \left(\frac{6}{14}\right)\left(\frac{14}{30}\right) = 0.2 \end{aligned}$$

(iii) P (neither solves the problem)

$$\begin{aligned} &= 1 - P(\text{either X or Y solve the problem}) \\ &= 1 - 0.7 = 0.30 \end{aligned}$$

Example 5.43: A market research firm is interested in surveying certain attitudes in a small community. There are 1250 households broken down according to income, ownership of a telephone and ownership of a T.V.

Survey Data from 1250 households

	Household with annual income Rs. 30,000 or less		Household with annual income above Rs. 30,000	
	Telephone Subscriber	No Telephone	Telephone Subscriber	No. Telephone
Own TV Set	270	200	180	100
No TV Set	180	100	120	100

- (i) What is the probability of obtaining a TV owner in drawing of households at random?
- (ii) If a household has annual income over Rs. 30,000 and is a telephone subscriber, what is the probability that he has a TV?
- (iii) What is the conditional probability of drawing a household that owns a TV, given that the household is a telephonic subscriber?
- (iv) Are the events 'ownership of a TV' and a 'telephone subscriber' statistically independent? Comment.

Solution:

	Household with annual income Rs.30, 000 or less		Household with annual income above Rs.30, 000		
	Telephone Subscriber	No Telephone	Telephone Subscriber	No Telephone	
Own TV Set	270	200	180	100	750
No TV Set	180	100	120	100	500
	450	300	300	200	1250

Let H_{TV} – household owning a TV

H_{Tel} – household owning a telephone subscription

H_{A3} – household with annual income above Rs.30, 000

Thus

(i) P (of obtaining a TV owner)

$$= P(H_{TV})$$

$$= \frac{750}{1250} = 0.6, \text{ from the above table}$$

$$(ii) P(H_{TV}/H_{A3} \cap H_{Tel})$$

$$= \frac{P(H_{TV} \cap H_{A3} \cap H_{Tel})}{P(H_{A3} \cap H_{Tel})}$$

$$= \frac{180}{300} = 0.6, \text{ from the above table}$$

$$(iii) P(H_{TV} / H_{Tel})$$

$$= \frac{P(H_{TV} \cap H_{Tel})}{P(H_{Tel})}$$

$$= \frac{450}{750} = 0.6$$

(iv) The event ownership of a 'TV' and 'Telephone subscriber' will be independent if

$$P(H_{TV} \cap H_{Tel}) = P(H_{TV})P(H_{Tel})$$

$$\text{L.H.S.} = P(H_{TV} \cap H_{Tel})$$

$$= \frac{450}{1250} = 0.36$$

$$\text{R.H.S.} = P(H_{TV})P(H_{Tel})$$

$$= \frac{750}{1250} \times \frac{750}{1250} = \frac{562500}{1562500} = 0.36$$

$$\text{Thus } P(H_{TV} \cap H_{Tel}) = P(H_{TV})P(H_{Tel})$$

Thus these two events are independent.

Example 5.44: A person decides to take a vacation. He has to choose between three places, Goa, Ladakh and Andamans. He is twice as likely to go to Andaman as to Goa and three times as likely to go to Goa as to Ladakh. What is the probability that he vacations in Goa?

Solution:

Let G: event that he vacations in Goa

L: event that he vacations in Ladakh

A: event that he vacations in Andaman

Then

$$P(A) = 2P(G)$$

$$3P(L) = P(G)$$

$$P(G) + P(L) + P(A) = 1$$

Let $x = P(G)$

Then

$$P(A) = 2x$$

$$P(L) = \frac{x}{3}$$

Thus

$$P(G) + P(A) + P(L) = 1$$

$$\Rightarrow x + 2x + \frac{x}{3} = 1$$

$$\Rightarrow 3x + \frac{x}{3} = 1$$

$$\Rightarrow 10x = 3$$

$$x = \frac{3}{10}$$

$$\Rightarrow x = 0.3$$

Thus probability that the person vacations in Goa is 0.3.

Example 5.45: The probability that a plane arrives at the airport before 11.00 am is 0.7. A passenger leaving from downtown takes a bus that arrives at the airport by 11.00 am with a probability of 0.6. Find the probability that

- (i) The plane and the passenger both arrive by 11.00 am.
- (ii) The plane arrives by 11.00 am and the passenger arrives after 11.00 am.

Solution:

Let A → The event that the plane arrives by 11.00 am.

B → The event that the passenger arrives by 11.00 am.

Then,

$$P(A) = 0.7 \text{ \& } P(B) = 0.6$$

Now

\bar{B} → The event that the passenger arrives after 11.00am.

$$\text{\& } P(\bar{B}) = 1 - P(B) = 1 - 0.6 = 0.4$$

(ii) P (plane & passenger both arrive by 11.00 am)

$$\begin{aligned} &= (A \cap B) \\ &= P(A) P(B) \quad (\because \text{Plane \& passenger travel independently}) \\ &= (0.7) (0.6) \\ &= 0.42 \end{aligned}$$

(iii) P (the plane arrives by 11.00 am and the passenger arrives after 11.00 am) is

$$\begin{aligned} &= P(A \cap \bar{B}) \\ &= P(A) \cdot P(\bar{B}) \\ &= (0.7) (0.4) \\ &= 0.28 \end{aligned}$$

Example 5.46: A MBA student applies for a job in two Firms – Firm I and Firm II. The probability of his being selected in Firm I is 0.7 and being rejected in Firm II is 0.5. The probability of at least one of his applications being rejected is 0.6. Find the probability that he would be selected in one of the firms?

Solution:

Let F_1 - Event that he is selected in Firm I

F_2 - Event that he is selected in Firm II

Given,

$$P(F_1) = 0.7 \Rightarrow P(\bar{F}_1) = 1 - 0.7 = 0.3$$

$$P(\bar{F}_2) = 0.5 \Rightarrow P(F_2) = 0.5$$

P (at least one of his application is rejected) = 0.6

Probability that he would be selected in one of the firms is:

$$P(F_1 \cup F_2) = P(F_1) + P(F_2) - P(F_1 \cap F_2)$$

Since, selection in the two firms are independent.

$$\begin{aligned} P(F_1 \cup F_2) &= P(F_1) + P(F_2) - P(F_1) P(F_2) \\ &= 0.7 + 0.5 - (0.7)(0.5) \\ &= 1.2 - 0.35 \\ &= 0.85 \end{aligned}$$

Example 5.47: In a city, three daily newspapers are published say A, B and C. 40% of the people of the city read newspaper A, 50% read B, 25% read C. 20% read both A & B, 15% read A and C and 8 % read B and C. 25% read all the 3 newspapers. What percentage of people in the city do not read any of the three newspapers?

Solution:

Let A - Event of reading newspaper A

B - Event of reading newspaper B

C - Event of reading newspaper C

Given the following probabilities

$$P(A) = 0.40 \quad P(AB) = 0.20 \quad P(ABC) = 0.25$$

$$P(B) = 0.50 \quad P(AC) = 0.15$$

$$P(C) = 0.25 \quad P(BC) = 0.08$$

By additive law for these events:

P (of reading at least one newspaper)

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC) \\ &= 0.40 + 0.50 + 0.25 - 0.20 - 0.15 - 0.08 + 0.25 \\ &= 1.15 - 0.43 + 0.25 \\ &= 0.97 \end{aligned}$$

And the required P (not reading any newspaper).

$$\begin{aligned} &= 1 - P(A \cup B \cup C) \\ &= 1 - 0.97 \\ &= 0.03 \end{aligned}$$

i.e. about 3% of the city population do not read any newspaper.

5.7 PROBABILITY DISTRIBUTION OF A RANDOM VARIABLE

5.7.1 Random Variable

A random variable assigns numerical values to the outcomes of a random or chance experiment. Thus, mathematically, it is a function defined on the outcomes of the sample space i.e. different values of the random variable are obtained by associating a real number with each element of the sample space.

For example:

(i) Suppose three coins are tossed simultaneously. The sample space is

$$\{HHH, HHT, HTH, THH, TTT, TTH, THT, HTT\}$$

Let X be a random variable denoting the number of heads.

The first element of the sample space is HHH i.e. three heads i.e. X assumes the value 3

$$X = 3$$

The second element contains 2 heads.

$$\text{i.e. } X = 2$$

The third and fourth element contains 2 heads

$$\text{i.e. } X = 2$$

The sixth, seventh and eighth contains 1 heads

$$X = 1$$

The fifth element contains 0 heads

$$\text{i.e. } X = 0$$

Thus, the values that the random variable X can assume are

$$X = 0, 1, 2, 3$$

Similarly we can define more random variables on the same sample space. For example, if

Y : The no of tails.

Or Z : There are at least two heads.

Then the values which Y and Z can assume are respectively.

$$Y: 0, 1, 2, 3$$

$$Z: 2, 3$$

Random variables are usually denoted by capital letters and the values they assume are denoted by small letters.

For Example $X = x_1$ would mean that the random variable X assume the value x_1 .

- (ii) Suppose a person gets four chances of throwing a dart at circular dart board. Let h denote the event that the person hits the centre and m denotes the event that the person misses.

Let X denote the number of hits of the person

Clearly, X is a random variable.

Now, the possibilities or sample space in all the four trails are:

$S = \{hhhh, hhhm, hhmh, hmhm, hmhm, mmmm, mmmh, mmhm, mhmh, hmmm\}$

i.e. 10 sample points.

The values that X can assume are 0, 1, 2, and 3, since in each sample points, there may be zero head, one head, two heads or three heads.

More example of random variables are:

- (i) Sales of a store in a day.
- (ii) No. of telephone calls received by an operator in a day.
- (iii) No. of vehicles at a traffic point
- (iv) No. of defective units produced in a production line
- (v) The amount of annual rainfall.
- (vi) Marks of students in an examination.
- (vii) Everyday price of a share

A random variable is either discrete or continuous depending on the values it assumes.

5.7.2 A discrete random variable is a variable that assumes a countable infinite number of values. i.e. the number of values that such a random variable assumes can be counted and these can be as many values as there are positive integers. Example (ii), (iii) and (iv) are examples of discrete random variables.

5.7.3 A continuous random variable is a random variable that can assume any value over an interval. Continuous variables can be measured to any desired degree of accuracy. The random variables in examples (v) and (vii) are continuous random variables.

5.7.4 Probability Distribution of a Random Variable

The probability distribution of a random variable is a listing of all possible values that a random variable can take along with their respective probabilities.

For example, in the coin tossing experiment used to describe random variables, the sample space was

$S = \{HHH, HHT, HTH, THH, TTT, TTH, THT, HTT\}$

The values that X can assume are

$X = 0, 1, 2, 3$

and

$$P(X = 0) = \frac{1}{8}$$

$$P(X = 1) = \frac{3}{8}$$

$$P(X = 2) = \frac{3}{8}$$

$$P(X = 3) = \frac{1}{8}$$

In tabular form, the above information can be represented as follows:

Table 5.2
Probability Distribution of Number of Heads

X	0	1	2	3	Total Probability
$P(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

It may be noted that the total probability is always unity.

The above listing of all possible values of a random variable along with the probabilities is known as the probability function of the random variable.

A probability distribution provides the possible values of random variable and their corresponding probabilities. It is a special case of the probability measure. Actually the theoretical counter part of a frequency distribution is called the probability distribution. A probability distribution of a random variable lists all possible values that it can take along with the corresponding probabilities.

There are basically two types of probability distributions depending upon the nature of the random variable viz.

- (i) Discrete Probability Distributions.
- (ii) Continuous Probability Distributions.

5.8 DISCRETE PROBABILITY DISTRIBUTIONS

Suppose X is a discrete random variable that assumes the values x_1, x_2, \dots, x_n along with respective probabilities $p(x_1), p(x_2), \dots, p(x_n)$.

Then, the probability function of the discrete random variable X, also called the probability mass function, is given as follows:

Table 5.3
Probability Distribution of a Discrete Random Variable

X	x_1	x_2	x_3	x_n
$P(X = x)$	$p(x_1)$	$p(x_2)$	$p(x_3)$	$p(x_n)$

$$\& \sum_{i=1}^n p(x_i) = 1$$

A probability mass function satisfies the following properties.

$$(i) 0 \leq p(x_i) \leq 1, \quad i = 1, 2, \dots, n$$

i.e. the probability that the random variable assumes the value x_i , $i = 1, 2, \dots, n$, lies between 0 and 1

$$(ii) \sum_{i=1}^n p(x_i) = 1$$

The sum of the probabilities is always 1

For example let X be a discrete random variable denoting the number on a dice. Then, X has a discrete probability distribution. The distribution is as follows:

Table 5.4
Probability Distribution of Number on a Dice

Value of X-the number on the dice	Probability
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
Total Probability	1

The Binomial distribution, Poisson distribution, Negative Binomial distribution are among the most well known discrete probability distributions.

5.8.1 Expected Value and Variance of a Discrete Probability Distribution

After establishing the probability distribution, the next step is to find out the characteristics of the distribution. In statistics, the characteristics often mean the expected value or mean and the variance of the distribution.

The expected value of a r. v. X . (denoted by) $E(X)$ is computed as a weighted average of the values of the random variable as follows:

$$E(X) = \sum_{i=1}^n x_i P(x_i)$$

$$= x_1 P(x_1) + x_2 P(x_2) + \dots + x_n P(x_n)$$

where

$P(x_i)$ = Probability of the random variable assuming value x_i

$E(X)$ = Expected value or mean or mathematical expectation of the random variable.

The variance of the probability distribution of a discrete random variable X denoted by $V(X)$ is

$$V(X) = E(X^2) - E^2(X)$$

$$\text{where } E(X^2) = \sum_{i=1}^n x_i^2 p(x_i)$$

The positive square root of the variance is called the standard deviation of the distribution.

$$\text{i.e. s.d. (X) = } \sqrt{V(X)}$$

Example 5.48: A company estimates the net profit on a new product to be launched shortly, to be 40,00,000 if it is successful, Rs. 15,00,000 if it is moderately successful and a loss of Rs. 10,00,000 if it fails in the market. The probabilities of the three different possibilities are respectively 0.20, 0.30 and 0.50. Find the expected value and the variance of the net profits.

Solution:

Let X denotes the net profit on the new product.

Then $X = 40, 15, -10$ (in Rs. '00, 000)

& $P(X) 0.20, 0.30, 0.50$

Thus,

$$\begin{aligned} E(X) &= 40 \times (0.20) + 15 \times (0.30) + (-10) \times (0.50) \\ &= 8 + 4.5 - 5 \\ &= 7.5 \end{aligned}$$

$$V(X) = E(X^2) - [E(X)]^2$$

$$\begin{aligned} E(X^2) &= \sum X^2 p(X) = 40^2 \times (0.20) + (15)^2 (0.30) + (-10)^2 (0.50) \\ &= 320 + 67.5 + 50 \\ &= 437.5 \end{aligned}$$

$$\text{Thus } V(X) = 437.5 - (7.5)^2 = 381.25$$

Example 5.49: There are 4 different choices available to a consumer who wants to buy a mobile phone of a particular company. The first model costs Rs. 9000, the second model costs Rs. 7800 and the third model Rs. 9800 and the fourth model costs Rs. 8600. The probabilities that the consumer would buy these models are $\frac{1}{3}$, $\frac{1}{6}$, $\frac{1}{4}$ and $\frac{1}{4}$ respectively.

The retailers commission on these models are 20%, 12%, 25% and 15% on the four sets respectively. Calculate the expected commission to be earned by the retailer.

Solution:

$$\text{Expected commission on the first model} = \frac{20}{100} \times 9000 = \text{Rs. } 1800$$

$$\text{Expected commission on the second model} = \frac{12}{100} \times 7800 = \text{Rs. } 936$$

$$\text{Expected commission on the third model} = \frac{25}{100} \times 9800 = \text{Rs. } 2450$$

$$\text{Expected commission on the fourth model} = \frac{15}{100} \times 8600 = \text{Rs. } 1290$$

Let $X \rightarrow$ Expected commission earned by the retailer

Then X can take the following values:

$$X = \text{Rs. } 1800, \text{Rs. } 936, \text{Rs. } 2450, \text{Rs. } 1290$$

and the probabilities associated with each of these commissions are respectively

$$P(X) = \frac{1}{3}, \frac{1}{6}, \frac{1}{4}, \frac{1}{4}.$$

Thus, in tabular form, the probability distribution of the retailers expected commission is

Probability Distribution of Retailers Expected Commission

X	1800	930	2450	1290
P (X = x)	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{4}$

Expected commission to be earned by the retailer:

$$\begin{aligned} E(X) &= \sum X P(X) = 1800 \times \frac{1}{3} + 930 \times \frac{1}{6} + 2450 \times \frac{1}{4} + 1290 \times \frac{1}{4} \\ &= 600 + 155 + 612.5 + 322.5 \\ &= \text{Rs. } 1690 \end{aligned}$$

Example 5.50: A man runs an ice cream parlor at a holiday resort. If the summer is mild, he can sell upto 3000 cups of ice cream, if it is hot, he can sell 5000 cups of ice cream. For any year, from past experience, the probability of a mild summer is $\frac{3}{7}$ and the probability of a hot summer is $\frac{4}{7}$.

A cup of ice-cream costs Rs. 3 and he sells it for Rs. 7. Find his expected profit.

Solution:

We first calculate the number of ice-creams that he expects to sell

Let $X \rightarrow$ no. of ice creams sold.

Then, we have the following probability distribution for X

	Mild Summer	Hot Summer
X	3000	5000
P(X = x)	$\frac{3}{7}$	$\frac{4}{7}$

$$E(X) = 3000 \times \frac{3}{7} + 5000 \times \frac{4}{7} = 4143 \text{ (approx) ice-creams}$$

Since, each ice-cream he sells for Rs. 7 and buys for Rs.3, his profit for each ice-cream sold = Rs. 4

$$\begin{aligned} \text{Thus, his expected profit} &= \text{Rs. } 4 \times 4143 \\ &= \text{Rs. } 16,572 \end{aligned}$$

Example 5.51: A property dealer pays 250 lakhs to bid on a contract. The probability of getting the contract is 0.2. If he gets the contract he makes a profit of Rs.10, 000 lakhs. If he does not get the contract, he forfeits the amount of Rs.250 lakhs. Find the dealers expected net profit.

Solution:

$$\begin{aligned} \text{Dealers expected net profit} &= 250 \times 0.8 + 10,000 \times 0.2 \\ &= 200 + 2000 = \text{Rs.}2200 \text{ lakhs} \end{aligned}$$

Example 5.52: A florist makes bouquets at the cost of Rs.10 each and sells them for Rs.30. The demand for bouquets has the following probability distribution.

Probability Distribution of Bouquets Sold

No. of Sold	0	1	2	3	4
p(x)	0.1	0.2	0.5	0.1	0.1

Find the expected number of bouquets to be sold. Also calculate profit on this number.

Solution:

$$\begin{aligned} \text{Expected number of bouquets to be sold} &= 0 (0.1) + 1 (0.2) + 2 (0.5) + 3 (0.1) + 4 (0.1) \\ &= 0.2 + 1 + 0.3 + 0.4 \\ &= 1.9 \\ &\cong 2 \end{aligned}$$

$$\begin{aligned} \text{Profit when 2 bouquets are sold} &= \text{Rs. } 30 \times 2 - \text{Rs.}10 \times 2 \\ &= \text{Rs. } 60 - \text{Rs.}20 = \text{Rs.}40 \end{aligned}$$

Example 5.53: Suppose there are three coins, which are tossed simultaneously. The possible results or outcomes are as follows. Find out the expected value for the number of tail in the experiment.

Probability Distribution of Number of Tails

Outcomes	TTT	TTH	HTT	THT	HHT	THH	HTH	HHH
No. of Tails	3	2	2	2	1	1	1	0

Solution:

Let number of tails be denoted by X

Then the expected values of tail E (X) is defined as

$$E(X) = \sum XP(X)$$

Calculation of Expected Value

No. of tail (X)	Probability P (x)	X P (x)
0	1/8	0
1	3/8	3/8
2	3/8	6/8
3	1/8	3/8
		$\frac{12}{8} = 1.5$

Thus, it can be said that on an average 1.5 tails can be expected.

Example 5.54: Consider the earlier example 5.53 and calculate σ^2 .

Solution:

Calculation of Variance

No. of tails	P (x)	E(x) = μ	(x - μ)	(x - μ) ²	(x - μ) ² P (x)
0	1/8	1.5	-1.5	2.25	0.28
1	3/8	1.5	-0.5	0.25	0.09
2	3/8	1.5	-0.5	0.25	0.09
3	1/8	1.5	1.5	2.25	0.28
					$\sum (x - \mu)^2 P(x) = .74$

So calculated variance is 0.74

We now discuss two well known discrete probability distributions viz. the Binomial Distribution and the Poisson Distribution.

5.8.2 The Binomial Distribution

Binomial distribution is one of the very popular and practically useful discrete probability distributions. Binomial distribution describes the possible number of times that a particular event will occur in a sequence of observations or trails.

The binomial distribution originates from the concept of a Bernoulli trial, named after the Swiss Mathematician Jacob Bernoulli (1654 - 1705). A Bernoulli trial is a trial which results in only two possible outcomes say, success or failure, yes or no, defective or non defective, hit or miss etc. The probability of a success in a Bernoulli trial is p and the probability of a failure in a Bernoulli trial is q . And $p + q = 1$.

Now, let X be a Bernoulli variable

$$\text{Then } X = \begin{cases} 0 & \text{with probability } q \\ 1 & \text{with probability } p \end{cases}$$

The random variable X is said to have a Bernoulli distribution.

Now consider n such Bernoulli trials and let X - no. of successes in n Bernoulli trials. The following assumptions are now made, which characterize a Binomial Distribution.

- (i) The trials are independent, i.e. outcome of one trial does not influence the outcome of any other trial.
- (ii) Each trial results in an outcome that can be classified as a success or a failure
- (iii) The probability of success i.e. p is known and remains the same in all trials.

$$p = P(\text{success}) \text{ \& } q = P(\text{failure}). \text{ And } p + q = 1.$$

The entire experiment consisting of n trials is a Binomial Experiment and our interest lies in the random variable X , where X = the number of successes in n trials.

In n trials the number of successes can be from 0 to n . Thus $X = 0, 1, 2, \dots, n$

$X = 0 \Rightarrow$ All failures or no success.

$X = 1 \Rightarrow$ 1 success. This is possible in $({}^n C_1)$ ways

$$n \text{ possible cases } \begin{cases} \text{SFF} \dots \text{F} \\ \text{FSF} \dots \text{F} \\ \cdot \\ \cdot \\ \text{FFF} \dots \text{S} \end{cases}$$

In general

$X = x \Rightarrow x$ successes and $(n - x)$ failures in n trials.

Thus, the probabilities are

$$P(X = 0) = ({}^n C_0) p^0 q^n$$

$$P(X = 1) = ({}^n C_1) p^1 q^{n-1}.$$

$$\text{In general } P(X = x) = {}^n C_x p^x q^{n-x}$$

Thus, in a binomial experiment with a constant probability p of success at each trial, the probability of x successes in n trials is given by

$$P(\mathbf{X} = \mathbf{x}) = {}^n\mathbf{C}_x p^x q^{n-x}, \mathbf{x} = 0, 1, 2, \dots, n$$

This is the probability mass function of the binomial distribution. The distribution can be summarized in the following table:

Table 5.5

Binomial Probabilities

\mathbf{X}	$\mathbf{0}$	$\mathbf{1}$	$\mathbf{..}$	\mathbf{x}	$\mathbf{..}$	\mathbf{n}
$P(\mathbf{X} = \mathbf{x})$	q^n	${}^n\mathbf{C}_1 p q^{n-1}$		${}^n\mathbf{C}_x p^x q^{n-x}$		p^n

Binomial probabilities for various values of n & p are given in the Binomial probability tables.

Example 5.55: Suppose a man fires 5 shots at a target. If he hits the target we can call it a success (p) and if he misses the target, it is considered a failure. Find the probability of two hits and all 5 hits.

Solution:

Let X – the number of hits

Then, X can assume the values 0, 1, 2, 3, 4, 5

By binomial probability

$$P(X = 2) = {}^5\mathbf{C}_2 p^2 q^3$$

$$P(\text{all 5 hits}) = {}^5\mathbf{C}_5 p^5$$

Example 5.56: A machine manufactures a certain kind of bolt in large numbers. The probability of finding 4 defective bolts in 10 is 0.111. Calculate the probability of getting 6 defective bolts.

Solution:

Let $X \Rightarrow$ Number of defective bolts

Here $n = 10$

$$P(X = 4) = 0.111 \quad \{\text{given}\}$$

$$\text{i.e. } ({}^{10}\mathbf{C}_4) p^4 q^6 = 0.111$$

From binomial tables $p = 0.6$. This may also be solved by putting $q = 1 - p$ in the above equation and then solving for p .

$$\text{Thus } q = 0.4$$

Now

$$\begin{aligned} P(X = 6) &= ({}^{10}\mathbf{C}_6) (0.1)^6 (0.4)^4 \\ &= 0.251 \quad (\text{from binomial tables or otherwise by solving}) \end{aligned}$$

Thus, probability of getting 6 defective bolts = 0.251

Example 5.57: A company received 8 bulbs packed in a box. The probability that a bulb is defective is 0.1. Find the probability of three defective bulbs.

Solution:

Let X-no. of defective bulbs

$$n = 8$$

p = Probability that a bulb is defective = 0.1

$$P(X = 3) = {}^8C_3 (0.1)^3 (0.9)^5 = 0.033$$

The probability of getting three defective bulbs = 0.033

Example 5.58: A radar complex consists of 8 units that operate independently. The probability that a unit detects an incoming missile is 0.90. What is the probability that an incoming missile will

- (i) not be detected by any unit.
- (ii) be detected by at most 4 units.

Solution:

Here

$$n = 8$$

p = probability that a unit detects an incoming missile = 0.90

X – no. of units that detect an incoming missile

- (i) $P(X = 0) = P(\text{none of the units detect an incoming missile})$
 $= {}^8C_0 (0.90)^0 (0.10)^8$
 $= (0.1)^8$
- (ii) $P(X \leq 4) = P(\text{at most 4 units detect an incoming missile})$
 $= 1 - [P(X = 5) + P(X = 6) + P(X = 7) + P(X = 8)]$
 $= 1 - [0.147 + 0.294 + 0.336 + 0.168]$
 $= 1 - [0.945] = 0.055$

Example 5.59: In a marketing survey, 50% of the shopkeepers of a particular region are ready to respond, then find out the probability that in a random sample at 10 shopkeeper, 6 have responded.

Solution:

Let the rate of response is the probability of success (p) then,

$$p = \frac{50}{100} = 0.5$$

Then probability of failure (q) i.e. a shopkeeper not responding is

$$q = (1-p) = (1 - 0.5) = 0.5$$

No. of trials: n = 10

Out of which the no. of success are: X = 6

Therefore

$$\begin{aligned} P(\text{six out of ten shopkeepers have responded}) &= P(X = 6) \\ &= {}^{10}C_6 (0.5)^6 (0.5)^{10-6} \\ &= \frac{10!}{6!(10-6)!} (0.5)^6 (0.5)^4 \\ &= 0.205 \text{ (from binomial table or an simplifying)} \end{aligned}$$

Example 5.60: 10 coins are tossed simultaneously. What is the probability of getting exactly 4 heads.

Solution:

Let p = probability of getting a head = $\frac{1}{2}$

q = probability of not getting a head = $\frac{1}{2}$

Let X denote the number of heads

Then the probability of getting 4 heads in a random throw of 10 coins simultaneously is given by

$$\begin{aligned} P(X = 4) &= {}^{10}C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{10-4} \\ &= {}^{10}C_4 \left(\frac{1}{2}\right)^{10} \\ &= 0.2051 \text{ (from binomial tables or by simplifying)} \end{aligned}$$

Example 5.61: In a factory that manufactures refrigerators, the quantity control department has adopted the following inspection plan for the days production. 10 refrigerators are selected randomly and if this sample contains less than 2 defective units, the entire production lot is accepted and shipped. What is the probability that a lot known to contain 20 percent defective items would be shipped?

Solution:

Let X be the number of defective units

$n = 10$

$p = P$ (a defective refrigerator in the lot) = 0.20

$q = 1 - p = 0.80$

The lot would be shipped if the no of defective units is less than 2

Thus P (shipping a lot with % defective 0.20) = P (number of defectives is less than 2)

$$\begin{aligned} \text{i.e. } P(X < 2) &= P(X = 0) + P(X = 1) \\ &= {}^{10}C_0 (0.20)^0 (0.80)^{10} + {}^{10}C_1 (0.20)^1 (0.80)^9 \\ &= 0.107 + 0.268 \\ &= 0.375 \end{aligned}$$

5.8.2.1 Mean of the Binomial Distribution

The mean (μ) or expected value X of a binomial distribution is as follows:

$$\begin{aligned} \mu = E(x) &= \sum_{x=0}^n xP(x) \\ &= \sum_{x=1}^n x \frac{n!}{x!(n-x)!} p^x q^{n-x} \end{aligned}$$

$$\begin{aligned}
&= \sum_{x=1}^n \frac{n!}{(x-1)!(n-x)!} p^n q^{n-x} \\
&= np \sum_{x=1}^{n-1} \frac{(n-1)!}{(x-1)!(n-x)!} p^{n-1} q^{n-x} \\
&= np(q+p)^{n-1} = np \qquad \because (q+p)^{n-1} = 1
\end{aligned}$$

Thus, **Mean of the Binomial Distribution = np**

5.8.2.2 Variance of the Binomial Distribution

For the Binomial Distribution,

$$\begin{aligned}
\text{Variance: } \sigma^2 &= \sum_{x=0}^n (x-\mu)^2 p(x) \\
&= \sum_{x=0}^n (x^2 + \mu^2 - 2\mu x) p(x) \\
&= \sum_{x=0}^n x^2 p(x) + \mu^2 \sum_{x=0}^n p(x) - 2\mu \sum_{x=0}^n x p(x) \\
&= \sum_{x=0}^n \{x(x-1) + x\} p(x) + \mu_x^2 \sum_{x=0}^n p(x) - 2\mu \sum_{x=0}^n x p(x). \text{ on simplifying,} \\
&= [n(n-1)p^2 + np] + \mu^2 - 2\mu\mu \qquad \text{as } \mu = np \\
&= n(n-1)p^2 + np - \mu^2 \\
&= (np)^2 - np^2 + np - (np)^2 = np(1-p) = npq
\end{aligned}$$

Thus $\sigma^2 = npq$ and $\sigma = \sqrt{npq}$ = standard deviation

It may be noted that, for the Binomial Distribution,

mean > variance.

5.8.2.3 Mode of the Binomial Distribution.

The mode of the Binomial Distribution depends on the value of $(n+1)p$ and may be unimodal or binomial

Case I: When $(n+1)p$ is an integer,

Then the binomial distribution is bimodal and the modes are:

$(n+1)p$ and $(n+1)p - 1$

Case II: When $(n + 1)p$ is not an integer.

In this case, the binomial distribution is unimodal and the mode is given by the integral part of the quantity $(n + 1)p$.

Remarks.

(1) The binomial distribution is completely characterized by the two parameters n and p .

(2) The total number of values that the random variable can assume are $n + 1$.

(3) The quantities ${}^n C_0, {}^n C_1, {}^n C_2, \dots, {}^n C_n$

are known as binomial coefficients and these co-efficient are symmetrical.

For various values of n , the values of these co-efficients can be obtained by using Pascal's Triangle.

Table 5.6
Pascal's Triangle

n	Binomial co-efficients						Sum of coefficients (2^n)		
1			1	1			$2^1 = 2$		
2			1	2	1		$2^2 = 4$		
3			1	3	3	1	$2^3 = 8$		
4			1	4	6	4	1	$2^4 = 16$	
5			1	5	10	10	5	1	$2^5 = 32$

Procedure to write the triangle

Step 1: In the first row, both coefficients are unity as ${}^n C_1 = {}^n C_0$

Step 2: In the second row, we write 1 in the beginning and the end, and the value of the middle co-efficient is obtained by adding the co-efficient of the first row.

Step 3: the other rows can be similarly written, by writing 1 in the beginning and end of each row and adding the two previous co-efficient to obtain the other middle trend, for example 3 in the 3rd row is obtained by adding 1 and 2 from the second row and so on.

3. Shape of the Binomial Distribution.

For small values of p ($p < 0.1$ usually) the Binomial Distribution is skewed to the right

As p increases, the skewness becomes less noticeable and for $p = 0.5$, the binomial distribution becomes symmetrical.

When p exceeds 0.5, the distribution gradually becomes skewed to the left.

Example 5.62: A factory has 15 machines, which may need adjustment from time to time. 5 of these machines are old and 10 are new. The probabilities of adjustment for old & new machines are $1/10$ and $1/20$ respectively.

Solution:

Let p_1 is the probability that an old machine needs adjustment = $1/10$

$$q_1 = 1 - p_1 = \frac{9}{10}$$

Let p_2 is the probability that a new machine needs adjustment = $1/20$

$$q_2 = 1 - p_2 = \frac{19}{20}$$

Then the probability that 3 old machines need adjustment

$$P_1(3) = {}^5C_3 \left(\frac{1}{10}\right)^3 \left(\frac{9}{10}\right)^{5-3} = {}^5C_3 \left(\frac{1}{10}\right)^3 \left(\frac{9}{10}\right)^2 = 0.0081 \text{ (from binomial tables or on simplifying)}$$

The probability that 1 new machine need adjustment

$$\begin{aligned} P_2(1) &= {}^{10}C_1 \left(\frac{1}{20}\right)^1 \left(\frac{19}{20}\right)^{10-1} \\ &= 0.3151 \text{ (from binomial tables or on simplifying)} \end{aligned}$$

Example 5.63: In a manufacturing industry, 10% of the items produced was found defective. Find out the probability that out of 15 items, which are selected as random, there are

- (i) Exactly 2 defectives
- (ii) At least 2 defectives

Also find out the mean and variance of the distribution.

Solution:

Let X = no. of defective items

p = probability of defective items = 0.10

n = 15

$$\begin{aligned} \text{(i) } P(X = 2) &= ({}^{15}C_2) (0.10)^2 (0.90)^{13} \\ &= 0.267 \end{aligned}$$

$$\begin{aligned} \text{(ii) } P(X \geq 2) &= 1 - P(X < 2) \\ &= 1 - [P(X = 0) + P(X = 1)] \\ &= 1 - [0.206 + 0.343] \\ &= 1 - [0.549] = 0.451 \end{aligned}$$

Example 5.64: The incidence of bird flu in India was such that chicken had 25% chance of suffering from it. What is the probability that out of 5 chicken, 3 or more will contract the disease?

Solution:

The probability of chicken suffering from the disease is

$$p = \frac{25}{100} = \frac{1}{4} = 0.25$$

The probability of chicken not suffering from bird flu is

$$q = 1 - P = 1 - \frac{1}{4} = \frac{3}{4} = 0.75$$

We have to find out the probability that 3 or more i.e. 4 or 5 out of 5 will contract the disease

$$\begin{aligned} P[X \geq 3] &= P[3] + P[4] + P[5] \\ &= {}^5C_3 \left(\frac{1}{4}\right)^3 \left(\frac{3}{4}\right)^2 + {}^5C_4 \left(\frac{1}{4}\right)^4 \left(\frac{3}{4}\right)^1 + {}^5C_5 \left(\frac{1}{4}\right)^5 \left(\frac{3}{4}\right)^0 \\ &= 0.0879 + 0.0146 + 0.0010 \text{ (from Tables)} \\ &= 0.1035 \end{aligned}$$

Example 5.65: (a) The mean of a binominal distribution is 4 and its standard demotion is $\sqrt{2}$. What are the values of n, p and q.

(b) The mean and variance of a Binomial distribution are 10 and 2 respectively. Find the probability that the variate takes the value 12.

Solution:

(a) Mean = $np = 4$

Variance = $npq = 2$

Now,

$$\frac{npq}{np} = q = \frac{2}{4} = \frac{1}{2}$$

$$\Rightarrow P = \frac{1}{2}$$

And $n = \frac{4}{p} = 4 \times 2 = 8$

Thus $n = 8$, $p = \frac{1}{2}$ and $q = \frac{1}{2}$.

(b) Mean = $np = 10$... (i)

Variance = $npq = 2$... (ii)

$$(ii) \div (i) \Rightarrow q = \frac{2}{10} = \frac{1}{5}$$

$$\Rightarrow p = \frac{4}{5}$$

$np = 10$ (= Mean)

$$\Rightarrow n = \frac{10}{p} = 10 \times \frac{5}{4} = 12.5 \approx 13 \text{ (approx.)}$$

$$\begin{aligned}\text{Now, } P(X = 12) &= {}^{13}C_{12} \left(\frac{4}{5}\right)^{12} \left(\frac{1}{5}\right)^1 \\ &= 0.179\end{aligned}$$

Example 5.66: A binomial random variable satisfies the relation $9P(x = 4) = P(x = 2)$ for $n = 6$. Find the value of the parameter p .

Solution:

Given,

$$\begin{aligned}9 P(x = 4) &= P(x = 2) \\ \Rightarrow 9 {}^6C_4 p^4 q^2 &= {}^6C_2 p^2 q^4 \\ \Rightarrow 9 p^2 &= q^2 \\ \Rightarrow 9 p^2 &= (1 - p)^2 \\ \Rightarrow 9 p^2 &= 1 - 2p + p^2 \\ \Rightarrow 8 p^2 + 2p - 1 &= 0 \\ \Rightarrow 8 p^2 + 4p - 2p - 1 &= 0 \\ \Rightarrow 4 p (2p + 1) - (2p + 1) &= 0 \\ \Rightarrow (4p - 1)(2p + 1) &= 0\end{aligned}$$

$$p = \frac{1}{4}, -\frac{1}{2}$$

Since, probability is always positive, we will consider

$$p = \frac{1}{4}$$

Example 5.67: In a binomial distribution consisting of 5 independent trials, the probabilities of 1 and 2 successes are 0.4096 and 0.2048 respectively. Find the probability of success.

Solution:

Number of Trials = $n = 5$

Let X be the number of successes.

$P(X = 1) = 0.4096$ (Given)

$$\Rightarrow {}^5C_1 p q^4 = 0.4096 \quad \dots (i)$$

and $P(X = 2) = 0.2048$ (Given)

$${}^5C_2 p^2 q^3 = 0.2048 \quad \dots (ii)$$

It may be observed that

$$P(X = 1) = 2 p(X = 2)$$

Using this condition and using (i) and (ii)

$${}^5C_1 p q^4 = 2 {}^5C_2 p^2 q^3$$

On simplifying,

$$\Rightarrow pq^4 = 4p^2q^3$$

$$\Rightarrow q = 4p$$

$$\Rightarrow 1 - p = 4p$$

$$\Rightarrow p = \frac{1}{5}$$

The probability of a success = $\frac{1}{5}$

Example 5.68: A company received 10 tubes packed in a box. The probability that a tube is defective is 0.1. Assuming independence, find

- The probability that there are 3 defective tubes
- The expected number of defective tubes.
- The standard deviation of the number of defective tubes.

Solution:

$$n = 10$$

$$p = \text{Probability that a tube is defective} = 0.1$$

Let X denote the number of defective tubes

$$(a) p(X = 3) = {}^{10}C_3(0.1)^3(0.9)^7 = 0.057$$

(b) The expected number of defective tubes is

$$E(X) = np = 10 \times 0.1 = 1$$

(c) The standard deviation of the number of defective tubes is

$$\begin{aligned} \text{s.d.}(X) &= \sqrt{npq} \\ &= 0.9487 \end{aligned}$$

Example 5.69: Suppose the company, in the above example received a consignment of 200 boxes of tubes, each box containing 10 tubes. If the probability that a tube is defective is 0.1, how many boxes can be expected to have 3 defective tubes.

Solution:

$$P(\text{Three defective tubes}) = 0.057$$

In a consignment of 200 boxes, the number of boxes expected to have 3 defective tubes is

$$\begin{aligned} &200 \times (0.057) \\ &= 11 \text{ boxes (approx).} \end{aligned}$$

5.8.2.4 Fitting a Binomial Distribution

Fitting a distribution to a given data implies determination of expected frequencies for different values of the random variable on the basis of the data.

To fit a binomial distribution to a given set of data, we use the following recurrence relation derived from the probability mass function of a binomial distribution.

$$\frac{P(x+1)}{P(x)} = \frac{{}^n C_{x+1} p^{x+1} q^{n-x-1}}{{}^n C_x p^x q^{n-x}}$$

Simplifying, we get the recurrence relation for probabilities of binomial distribution

$$P(x+1) = \left\{ \frac{n-x}{x+1} \right\} \frac{p}{q} P(x), \quad x = 0, 1, 2, \dots, n-1.$$

The steps of fitting a binomial distribution are

Step 1: Calculate $p(0) = q^n$

If p and q are not known, we calculate the mean using the formula $\bar{x} = \frac{1}{N} \sum_{i=1}^n x_i f_i$ and set it equal to np

$$\text{Thus } \hat{p} = \frac{\bar{x}}{n} \text{ and } \hat{q} = 1 - \hat{p}$$

Step 2: Once $p(0)$ is calculated, the rest of the probabilities are calculated by using the recurrence formula

$$\text{Thus } P(1) = \left\{ \frac{n-1}{1+1} \right\} \frac{p}{q} P(0) \text{ when } x = 0$$

$$P(2) = \left\{ \frac{n-2}{2+1} \right\} \frac{p}{q} P(1) \text{ and so on}$$

Step 3: The expected frequencies are finally obtained by the formula $NP(x)$ for all values of x .

Example 5.70: Fit a binomial distribution to the following data:

Frequency Distribution

x	0	1	2	3	4
f(x)	28	62	28	12	46

Solution:

Step 1: We have to first calculate p and q

$$n = 4, \quad \sum f(x) = 176, \quad \sum x f(x) = 338$$

$$\text{Mean} = np = \frac{1}{N} \sum x f(x) = 1.92$$

$$\Rightarrow p = \frac{1.92}{4} = 0.48$$

$$\Rightarrow q = 1 - 0.48 = 0.52 \quad N = 176$$

$$\begin{aligned}\text{Thus } p(0) &= (0.52)^4 \\ &= 0.073\end{aligned}$$

$$\text{and } \frac{p}{q} = 0.923$$

Step 2: We now calculate the rest of the probabilities by using the recurrence formula

$$p(x+1) = \left\{ \frac{n-x}{x+1} \right\} \frac{p}{q} p(x), \quad x = 0, 1, 2, 3$$

$$x = 0 \Rightarrow p(1) = \frac{n-0}{0+1} \frac{p}{q} p(0) = 4(0.92)(0.073) = 0.268$$

$$x = 1 \Rightarrow p(2) = \frac{n-1}{2} \frac{p}{q} p(1) = \frac{3}{2}(0.268)(0.92) = 0.369$$

$$x = 2 \Rightarrow p(3) = \frac{n-2}{3} \frac{p}{q} p(2) = (0.67)(0.92)(0.369) = 0.23$$

$$x = 3 \Rightarrow p(4) = \frac{n-3}{4} \frac{p}{q} p(3) = (0.25)(0.92)(0.23) = 0.0529$$

Step 3: The frequencies are obtained by the formula $NP(x)$

$$f(0) = 176 \times (0.077) \cong 14$$

$$f(1) = 176 \times (0.27) \cong 48$$

$$f(2) = 176 \times (0.37) \cong 65$$

$$f(3) = 176 \times (0.23) \cong 40$$

$$f(4) = 176 \times (0.053) \cong 9$$

Fitted Binomial Distribution

x	0	1	2	3	4
p (x)	0.077	0.27	0.37	0.23	0.053
F (x)	14	48	65	40	9

which gives the final fitted binomial distribution

5.8.3 Poisson Distribution

Poisson Distribution, derived by the noted mathematician Simon D. Poisson in 1837 is also one of the important discrete probability distributions. It expresses the probability of a number of events occurring within a given time interval. Poisson distribution is widely used in business management. A few examples of situations that can be analysed by Poisson distribution are as follows:

- (i) The occurrence of accidents in a factory in a week.
- (ii) The number of phone calls received in an hour.
- (iii) Errors in typing per page.
- (iv) The number of insurance claims per year.

It is possible to derive the Poisson distribution as a limiting case of the binomial distribution where p the probability of success in a trial is very small, n is large and $np = \lambda$ is a infinite constant number. It follows the following distribution:

Table 5.7
Poisson Probabilities

X	0	1	2	n
Probability	$e^{-\lambda}$	$\frac{e^{-\lambda}\lambda}{1!}$	$\frac{e^{-\lambda}\lambda^2}{2!}$	$\frac{e^{-\lambda}\lambda^n}{n!}$

The probability mass function of the random variable X represents the number of times the event occurs in a given interval of time. This mass function can be written as

$$P(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0, 1, 2, \dots$$

λ is said to be the parameter of the distribution. Mathematically $X \sim P(\lambda)$ is the notation used for a random variable X following Poisson Distribution with parameter λ .

5.8.3.1 Mean of the Poisson Distribution

Let X be a Poisson variable with mean λ i.e. $X \sim p(\lambda)$.

$$\begin{aligned} \mu &= E(X) = \sum_{x=0}^{\infty} xP(x) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} \\ &= 0 + \lambda e^{-\lambda} + \lambda^2 e^{-\lambda} + \frac{\lambda^3 e^{-\lambda}}{2!} + \frac{\lambda^4 e^{-\lambda}}{3!} + \dots \\ &= \lambda e^{-\lambda} \left[1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots \right] \\ &= \lambda e^{-\lambda} e^{\lambda} = \lambda \end{aligned}$$

Thus, mean of Poisson distribution = λ

5.8.3.2 Variance of the Poisson Distribution

$$\begin{aligned} \text{Variance} &= \sigma^2 = E(X^2) - [E(X)]^2 \\ &= E(X^2) - \lambda^2 \end{aligned}$$

$$= \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda} \lambda^x}{x!} - \lambda^2$$

Now,

$$\sum_{x=0}^{\infty} \frac{x^2 e^{-\lambda} \lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{x(x-1) + x}{x!} \lambda^x$$

Thus,

$$\begin{aligned} \sigma^2 &= \lambda^2 e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} - \lambda^2 \\ &= \lambda^2 e^{-\lambda} \cdot e^{\lambda} + \lambda e^{-\lambda} e^{\lambda} - \lambda^2 \\ &= \lambda^2 + \lambda - \lambda^2 = \lambda \end{aligned}$$

Variance of Poisson distribution = λ

Thus, mean of the Poisson distribution = Variance of the Poisson distribution.

5.8.3.3 Mode of the Poisson distribution

As in the case of the Binomial distribution, the mode of Poisson distribution may be unimodal or bimodal depending on the value of the mean λ

Case I: If λ is an integer, the Poisson distribution is bimodal and the modes are given by λ and $\lambda - 1$.

Case II : If λ is not an integer, the Poisson distribution is unimodal and is given by the integral part of λ .

5.8.3.4 Poisson Approximation to Binomial

The Poisson distribution may be used as an approximation to Binomial under the following conditions:

- (i) When n , the number of trials become large i.e. $n \geq 20$
- (ii) p , the probability of success ≤ 0.05 i.e. p is very small.

Thus, under the above conditions, if $X \sim B(n, p)$ then X can be approximated as a Poisson variable with mean $\lambda = np$ i.e. $X \sim P(np)$

5.8.3.5 An Application of the Poisson Distribution.

The most common application of Poisson Distribution is in queuing theory problems. Suppose, customers arrive at a shop during a time interval, independent of each other. Then the distribution of the number of arrivals can be described with the help of the Poisson distribution. Such a process is called a Poisson process and the chief characteristics of this process are as follows:

- (i) The number of occurrences in an interval of time is independent of the number of occurrences in another interval.
- (ii) The expected number of occurrences in an interval is constant.

- (iii) It is possible to identify an interval of such short duration that the probability of occurrence of more than one event in such an interval is zero.

Remarks

- (i) The Poisson distribution is completely characterized by a single parameter i.e. the mean.
 (ii) The Poisson distribution is a positively skewed distribution.
 (iii) It is normally applied in situations when the number of trials is large and the probability of a success in a trial is very small.

Example 5.71: A manufacturer of pins observed that 4% of his product is defective. If the manufacturer sells 200 pins and gives guarantee that not more than 10 pins will be defective. What is the probability of failing guarantee?

Solution:

$$n = 200$$

Let X = Number of defective pins

Let p = the probability of defective pin = 4%

$$\text{Mean defective pins} = np = 200 \times 0.04 = 8$$

The manufacturer will fail to meet the guarantee if the number of defective pin is observed to be more than 10.

Thus, here we have to calculate $P(X > 10)$

$$\begin{aligned} &= 1 - P(X \leq 10) = \sum_{x=0}^{10} \frac{e^{-8} 8^x}{x!} \\ &= 1 - 0.8159 = 0.1841 \end{aligned}$$

Example 5.72: In a book of 500 pages, 300 typing errors were observed. Assuming Poisson distribution, find the probability that 1 page will contain no error?

Solution:

$$\text{The average error} = \lambda = \frac{300}{500} \times 1.0 = 0.6$$

The probability of no error in a page is calculated as

$$P(x=0) = \frac{e^{-0.6} 0^0}{0!} = e^{-0.6} = 0.5488$$

Example 5.73: If 2% of halogen lamps manufactured by a company are defective, find the probability that in a sample of 200 bulbs,

- (i) less than 2 bulbs are defective.
 (ii) more than 3 bulbs are defective.

Solution: We may use the Poisson Approximation to the Binomial Distribution

$$\text{Mean defective bulbs} = \frac{2}{100} \times 200 = 4$$

Let $X \rightarrow$ The number of defective bulbs.

$$(i) P(X < 2) = p(X = 0) + p(X = 1) + p(X = 2)$$

$$= \frac{e^{-4}4^0}{0!} + \frac{e^{-4}4^1}{1!} + \frac{e^{-4}4^2}{2!}$$

$$= 0.0183 + 0.0733 + 0.1465$$

$$= 0.2381$$

$$(ii) P(X > 3) = 1 - P(X \leq 3) = 1 - \{P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)\}$$

$$= 1 - \{0.2381 + 0.1954\}$$

$$= 1 - 0.4335$$

$$= 0.5665$$

Example 5.74: If X is a Poisson variate such that $P(X) = P(X + 1)$, find mean and standard deviation of X .

Solution:

Given condition, $P(X) = P(X + 1)$

$$\Rightarrow \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-\lambda}\lambda^{x+1}}{(x+1)!}$$

$$\Rightarrow \frac{(x+1)x!}{x!} = \lambda$$

$$\Rightarrow \lambda = x + 1$$

Thus, mean = $x + 1$

Standard deviation = $\sqrt{x+1}$

Example 5.75: If X is a Poisson variate such that

$P(X = 2) = 9P(X = 4) + 90P(X = 6)$, find the mean and variance of X .

Solution:

Let $X \sim P(m)$

Given: $P(X = 2) = 9 P(X = 4) + 90 P(X = 6)$

$$\Rightarrow \frac{e^{-m} m^2}{2!} = 9 \frac{e^{-m} m^4}{4!} + 90 \frac{e^{-m} m^6}{6!}$$

$$\Rightarrow \frac{1}{2!} = \frac{9 m^2}{24} + \frac{90 m^4}{720}$$

$$\Rightarrow m^4 + 3m^2 - 4 = 0$$

On simplifying, $m = 1$.

Thus, mean = 1, Variance = 1.

Example 5.76: A retailer purchases torches in lots of 400. He inspects 20 torches from each lot and accepts the lot if the number of defectives in each lot is not more than 2. Suppose a lot containing 30 defective torches is received by the retailer, what is the probability that it would be rejected?

Solution:

This problem can be solved by using the Poisson approximation of Binomial distribution

Here $n = 400$

and $p =$ Probability of finding a defective torch in the lot

$$= \frac{30}{400} = 0.075$$

Since n is reasonably large, p is small and $np = 400 \times (0.075) = 30$ which is finite. We can use the Poisson distribution instead of the Binomial distribution.

Let X be the number of defectives

Then $X \sim P(30)$

The lot is rejected if the number of defectives exceed 2,

$$\begin{aligned} P(X > 2) &= 1 - P(X \leq 2) \\ &= 1 - \{P(X = 0) + P(X = 1) + P(X = 2)\} \\ &= 1 - \{0.2230 + 0.3347 + 0.2510\} \\ &= 1 - 0.8087 \\ &= 0.1913 \end{aligned}$$

which is the required probability

Example 5.77: Between 2 and 4 pm, the number of phone calls coming into the switch board of a company in 300. Find the probability that during one particular minute there will be (i) no phone call at all (ii) exactly 3 calls.

Solution:

Since the number of calls coming between 2 & 4 pm is 300, the mean numbers of calls per minute $\frac{300}{120} = 2.5$ calls per minute

Let X denote the number of calls per minute

Then X has a Poisson distribution with mean

$$\lambda = 2.5$$

i.e. $X \sim P(2.5)$

(i) Probability that in a minute there would be no phone call

$$\begin{aligned} P(X = 0) &= \frac{e^{-2.5} 2.5^0}{0!} \\ &= 0.08208 \end{aligned}$$

(ii) Probability that these would be exactly 3 calls.

$$\begin{aligned} P(X = 3) &= \frac{e^{-2.5} (2.5)^3}{3!} \\ &= 0.21375 \end{aligned}$$

5.8.3.6 Fitting of Poisson Distribution

The steps in fitting a Poisson distribution are similar to that of the Binomial distribution. The recurrence relation in this case is calculated as follows:

$$\frac{p(x+1)}{p(x)} = \frac{e^{-\lambda} \lambda^{x+1}}{e^{-\lambda} \lambda^x} \cdot \frac{x!}{(x+1)!}, \quad x = 0, 1, 2, \dots$$

Simplifying, the recurrence relation for probabilities of a Poisson Distribution are

$$\frac{p(x+1)}{p(x)} = \frac{\lambda}{x+1}$$

$$\Rightarrow p(x+1) = \frac{\lambda}{x+1} p(x), \quad x = 0, 1, 2, \dots$$

The steps can be summarized as follows:

Step 1: Calculate $P(0) = e^{-\lambda}$

If λ is not known, we calculate the mean and equate it to λ .

Step 2: Calculate the rest of probabilities using the above recurrence relation

For example,

$$p(1) = \frac{\lambda}{2} p(0)$$

$$p(2) = \frac{\lambda}{3} p(1) \dots \text{and so on}$$

Step 3: Calculate the frequencies using the formula

$$F(x) = Np(x)$$

$$\text{Thus, } F(0) = Np(0)$$

$$F(1) = Np(1)$$

.

.

.

and so on, where N is the sum of all the frequencies.

Example 5.78: The following table shows the number of customers returning a certain product because of defects. The data from 100 stores:

Distribution of the number of returns of a defective product

No. of returns	0	1	2	3	4	5	6
No. of stores	4	14	23	23	18	9	9

Fit a Poisson distribution.

Solution:

Since λ is unknown, we first calculate it.

Step 1: $\lambda = \frac{1}{N} \sum xf(x)$, $N = \sum f(x)$

Calculation of mean

No. of returns (x)	No. of stores f (x)	x f(x)
0	4	4
1	14	14
2	23	46
3	23	69
4	18	72
5	9	45
6	9	54
	100	304

$$\lambda = \frac{304}{100} = 3.04 \cong 3$$

$$p(0) = e^{-3} = 0.0498 \text{ (from Poisson tables)}$$

Step 2:

Using the recurrence relation for probabilities of Poisson Distribution

$$\text{Let } x = 0 \Rightarrow p(1) = \frac{\lambda}{1} p(0) = 3 = 0.15$$

$$x = 1 \Rightarrow p(2) = \frac{\lambda}{2} p(1) = 0.225$$

$$x = 2 \Rightarrow p(3) = \frac{\lambda}{3} p(2) = 0.225$$

$$x = 3 \Rightarrow p(4) = \frac{3}{4} p(3) = (0.75)(0.225) = 0.169$$

$$x = 4 \Rightarrow p(5) = \frac{3}{5} p(4) = (0.6)(0.169) = 0.1024$$

$$x = 5 \Rightarrow p(6) = \frac{3}{6} p(5) = (0.5)(0.1024) = 0.081$$

Step 3: The expected frequencies are calculated in the following table:

Calculation of expected frequencies

(x)	p (x)	F (x) = Np (x)
0	0.04	4
1	0.15	15
2	0.225	23
3	0.225	23
4	0.169	17
5	0.101	10
6	0.081	8
	100	

The fitted Poisson distribution is

Fitted Poisson Distribution

x	0	1	2	3	4	5	6
F (x)	4	15	23	23	17	10	8

5.9 CONTINUOUS PROBABILITY DISTRIBUTION

Continuous probability distribution is defined for an infinite number of points over a continuous interval. The normal distribution described in the next section is the most popular and widely used continuous distribution.

The probability function of a continuous random variable is called a probability density function. Specifically, if there are two numbers a and b , then the probability that the random variable X assumes a value between a and b , is written as

$$P(a < X < b)$$

This probability represents the area under the probability density curve (of the probability density function of X) between a and b .

It may be noted that, for a continuous random variable, the probability that the random variable assumes a particular value is zero i.e.

$$P(X = a) = 0, \text{ for any } a.$$

Thus, for a continuous random variable, the following expressions are equivalent.

$$P(a < x < b) = P(a \leq x \leq b) = P(a \leq x < b) = P(a < x \leq b)$$

5.9.1 Normal distribution

Normal distribution also known as Gaussian distribution is an extremely important continuous probability distribution.

A large number of natural phenomenon may be represented by the normal distribution. Most of the pioneering work in Gaussian distribution may be attributed to the German Mathematician Karl Friedrich Gauss.

This distribution has two parameters, the mean and the standard deviation. The population mean is denoted by μ and the population standard deviation by σ . The probability function of normal distribution with mean μ and standard deviation as follows

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty \leq x \leq \infty, \quad -\infty < \mu < \infty, \quad \sigma \geq 0$$

π, e are the constant terms & assume the following values

$$\pi = 3.1416$$

$$e = 2.7183$$

Symbolically, a normal distribution of the random variable X with mean μ and standard deviation σ is expressed as

$$X \sim N(\mu, \sigma)$$

The parameters μ and σ completely characterize the distribution. The graph of normal distribution is a bell shaped curve, popularly known as the normal curve and is shown in the following figure. This curve is symmetric about the mean of the distribution.

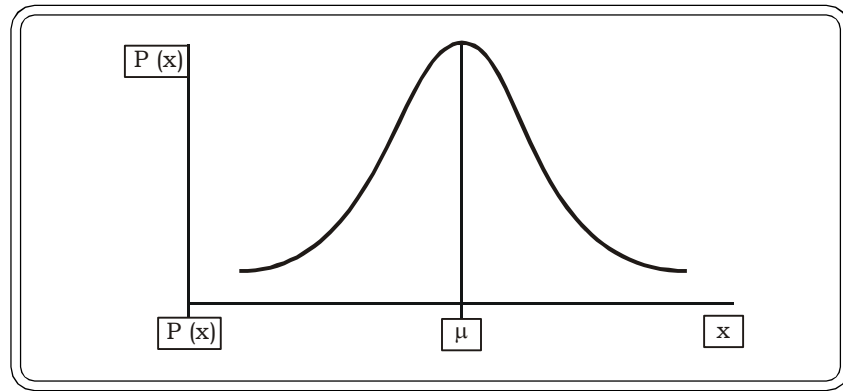


Figure 5.13

Graph of a Normal Distribution

Standard Normal Variate

A normal random variable say X can be transformed to a standard normal variate by subtracting the mean (μ) and dividing by the standard deviation (σ). If we denote the standardized variate as Z , then

$$Z = \frac{X - \mu}{\sigma}$$

The probability density function of the standard normal variate Z is as follows:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

$$E(Z) = 0 \text{ and } V(Z) = 1. \text{ Thus, } Z \sim N(0, 1)$$

Any value of the standardized variate indicates the distance, expressed as a multiple of the standard deviation, that the value lies away from the mean. For example $z = \pm 1$, implies that the value x lies 1 standard deviation above and below the mean.

5.9.2 Characteristics of Normal Distribution

Some characteristics of the normal distribution are as follows:

- (i) The normal distribution is symmetric about the mean. Thus, for a normal distribution, $\beta_1 = 0$, and $P(X > \mu) = 0.50 = P(X < \mu)$.
- (ii) For the normal distribution, the mean, median and mode coincide i.e.

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

- (iii) The tails of the curve extend indefinitely in both directions from the center such that the curve is asymptotic to the X -axis. Thus, though the curve tails get closer and closer to the X -axis they never really touch the X -axis or mathematically, they connect with the X -axis at infinity.
- (iv) The total area of the curve gives the total probability of the random variable taking values between $-\infty$ and ∞ . Thus

$$P(-\infty < x < \infty) = \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sqrt{2\pi}\sigma} dx = 1$$

- (v) The mean μ determines the center of the curve and the standard deviation σ determines its flatness.

The following figure shows normal curves having different means, but same variance.

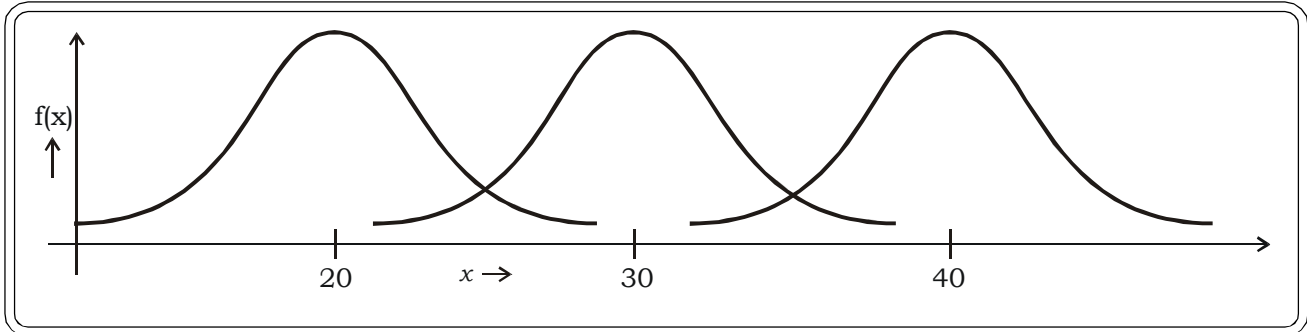


Figure 5.14

Normal Curves with Different Means and same Variance

The figure below shows normal curves with same mean but different variances.

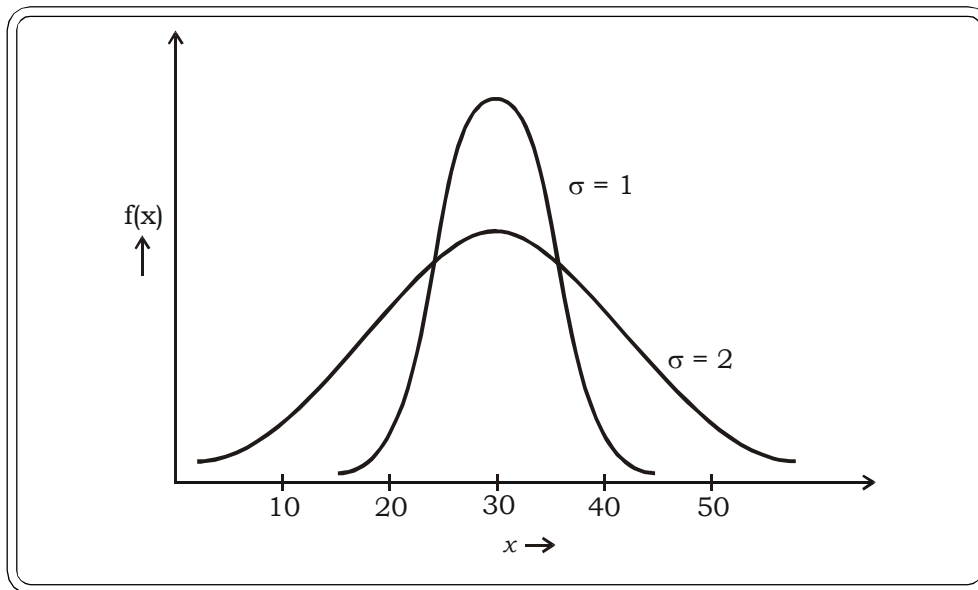


Figure 5.15

Normal Curves with same Mean and Different Variances

(vi) Additive Property of the Normal Distribution

Let, $X_1 \sim N(\mu_1, \sigma_1)$ and $X_2 \sim N(\mu_2, \sigma_2)$

and suppose X_1 and X_2 are independent.

then let $Z = aX_1 + bX_2$ i.e. Z is a linear combination of X_1 and X_2 .

$$\text{Then, } Z \sim N(a_1\mu_1 + b\mu_2, a^2\sigma_1^2 + b\sigma_2^2)$$

When $a = 1, b = 1$

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

This result can be generalized to n independent normal variables.

(viii) **Area Property of normal variables.**

If $X \sim N(\mu, \sigma^2)$, Then

$$(a) P(\mu - \sigma < X < \mu + \sigma) = 0.6826$$

$$(b) P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9544$$

$$(c) P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$$

The three probabilities given above can be interpreted as follows:-

For a normal distribution

- 68% of the observations would be between $\mu - \sigma$ and $\mu + \sigma$ or within ± 1 standard deviation from the mean.
- About 95% of the observations would be between $\mu - 2\sigma$ and $\mu + 2\sigma$ or within ± 2 standard deviations from the mean.
- About 99% of the observations would be between $\mu - 3\sigma$ and $\mu + 3\sigma$ or within ± 3 standard deviations from the mean.

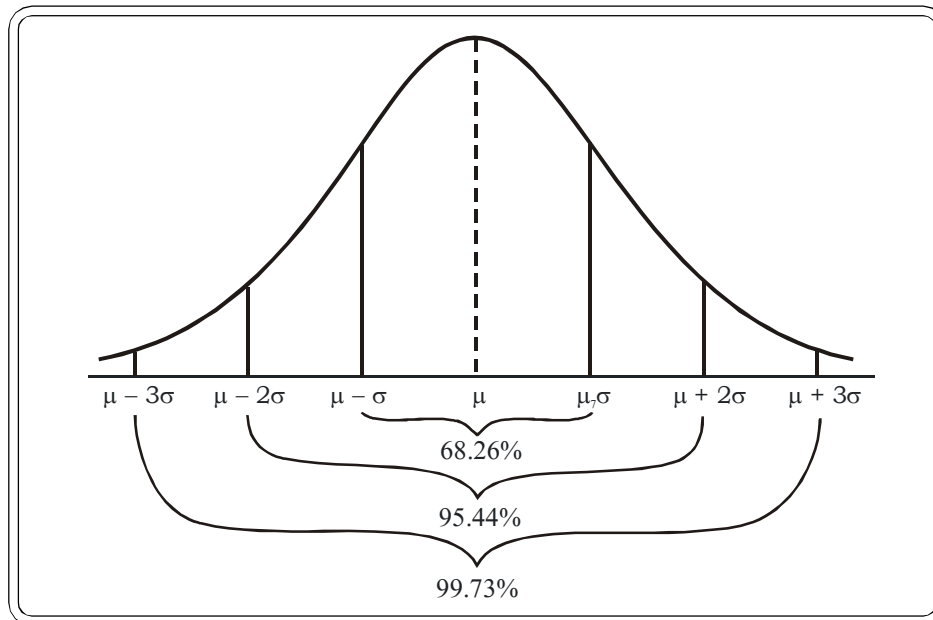


Figure 5.16

Area Property of Normal Distribution

Example 5.79: Let X be a random variable which follows a normal distribution with mean 25 and variance 4. Find:

- (i) $P(23 < X < 27)$
- (ii) $P(X > 26)$
- (iii) $P(X < 20)$

Solution:

- (i) $P(23 < X < 27)$

Converting to the standard normal or Z - scale i.e. $z = \frac{X - \mu}{\sigma}$, Here $\mu = 25$, $\sigma = 2$

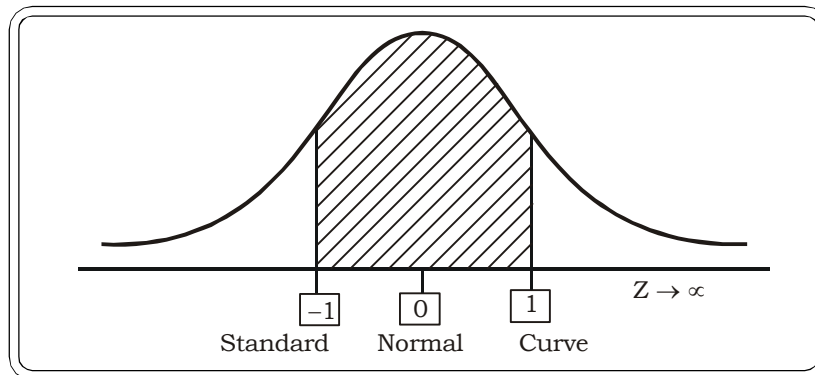
$$= P\left(\frac{23 - 25}{2} < \frac{X - 25}{2} < \frac{27 - 25}{2}\right)$$

$$= P(-1 < Z < 1)$$

Z follows a standard normal distribution with mean 0 and variance 1.

The normal tables give areas of the normal curve from 0 to a certain positive number (≤ 3)

In this example we want the area from -1 to 1



The standard normal tables will give us area from 0 to 1. Since the curve is symmetric this area will be equal to the area from -1 to 0.

Thus

$$\begin{aligned} P(-1 < Z < 1) &= 2P(0 < Z < 1) \\ &= 2(0.3413) \text{ from tables} \\ &= 0.6826 \end{aligned}$$

This can also be simply derived from the 1st area property by putting $\mu = 0$ and $\sigma = 1$ in

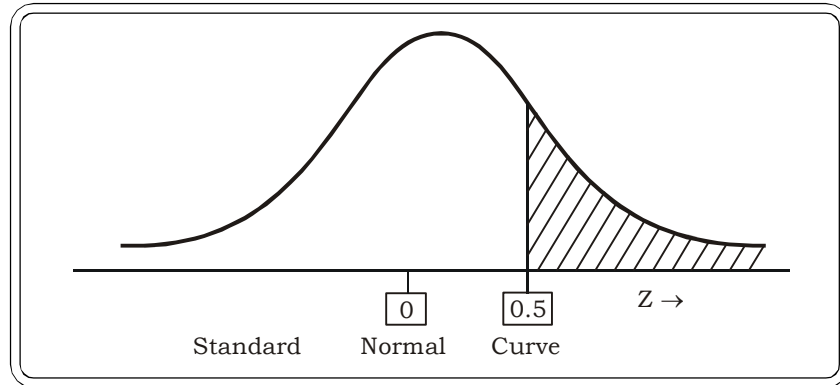
$$P(\mu - \sigma < x < \mu + \sigma) = 0.6826$$

- (ii) $P(X > 26)$

Converting to the Z - scale

$$= P\left(Z > \frac{26-25}{2}\right)$$

$$= P(Z > 0.5)$$



$$= 0.5 - P(0 < Z < 0.5)$$

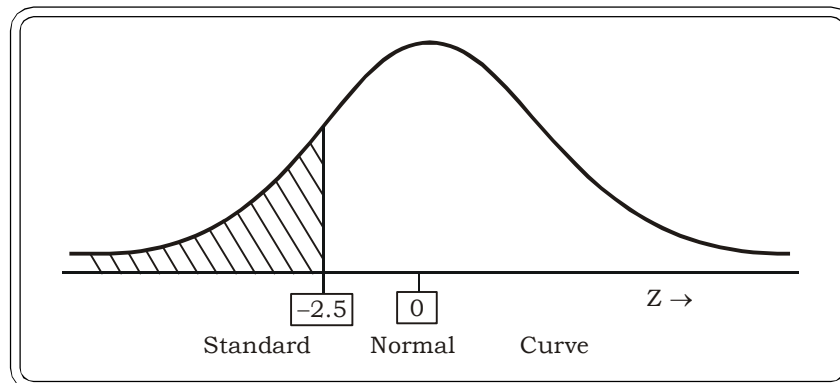
$$= 0.5 - 0.1915$$

$$= 0.3085$$

(iii) $P(X < 20)$

$$= P\left(Z < \frac{20-25}{2}\right) \text{ \{converting to the } z \text{- scale\}}$$

$$= P(Z < -2.5)$$



$$= P(Z > 2.5). \text{ Since the normal curve is symmetric.}$$

$$= 0.5 - P(0 < Z < 2.5)$$

$$= 0.5 - 0.4938$$

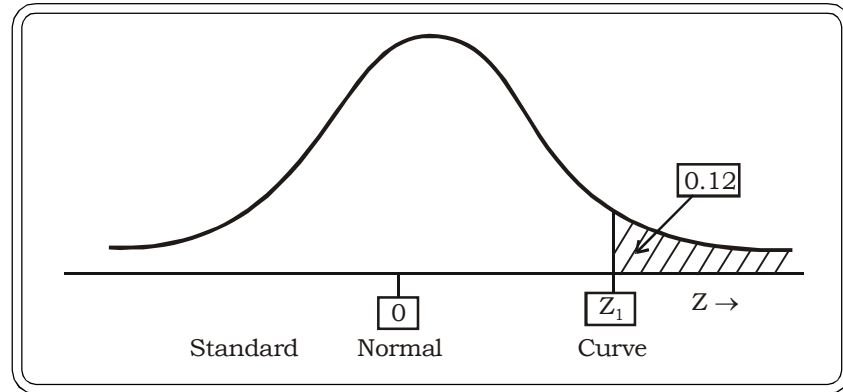
$$= 0.0062$$

Example 5.80: Consider a normal distribution with standard deviation 10. The probability that a value selected from this distribution exceeds 105 is 0.12. Find the mean of this distribution.

Solution: Given

$$P(X > 105) = 0.12$$

$$\Rightarrow P\left(Z > \frac{105 - \mu}{10}\right) = 0.12$$



$$= P(Z > Z_1) = 0.12$$

From tables

$$P(0 < Z < 1.17) = 0.38$$

$$\Rightarrow Z_1 = 1.17$$

$$\Rightarrow \frac{105 - \mu}{10} = 1.17$$

$$\begin{aligned} \Rightarrow \mu &= 105 - 11.7 \\ &= 93.3 \end{aligned}$$

Mean of the distribution = 93.3

Example 5.81: Average daily sales at 800 retail stores, selling baby soap was Rs.150 thousand with a standard deviation of 15 thousand. If the sales of baby soap are assumed to follow a normal distribution, find how many retail stores have sales between

- (i) Rs.130 thousand and 145 Thousand
- (ii) More than Rs.160 thousand
- (iii) Less than 120 thousand

Solution:

Let X represent the daily sales of baby soap of 800 retail stores.

Then X follows a normal distribution with

Mean Sales = Rs.150 thousand

and Standard Deviation = Rs.15 thousand

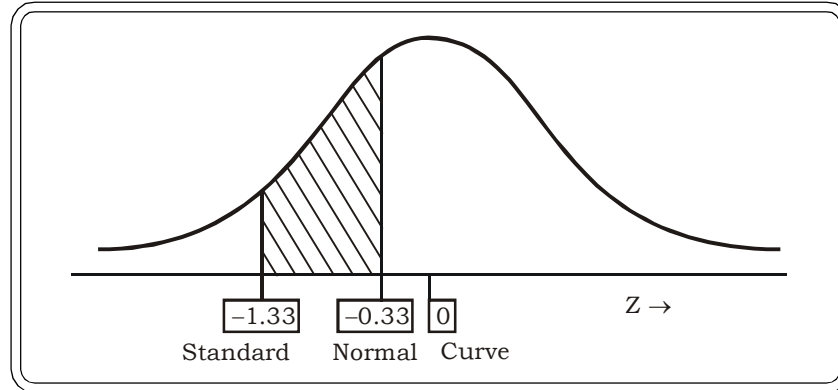
- (i) We need to evaluate no. of stores having sales between Rs.130 thousand and 145 thousand. We first find the

$$P(130 < X < 145)$$

Converting to the standard Z - scale

$$= P\left(\frac{130-150}{15} < \frac{X-150}{15} < \frac{145-150}{15}\right)$$

$$= P(-1.33 < Z < -0.33)$$



$$= P(0.33 < Z < 1.33) \quad \{\because \text{the curve is symmetric}\}$$

$$= P(0 < Z < 1.33) - P(0 < Z < 0.33)$$

$$= 0.4082 - 0.1293$$

$$= 0.2789$$

The no. of stores having sales between Rs.130 thousand and Rs.145 thousand

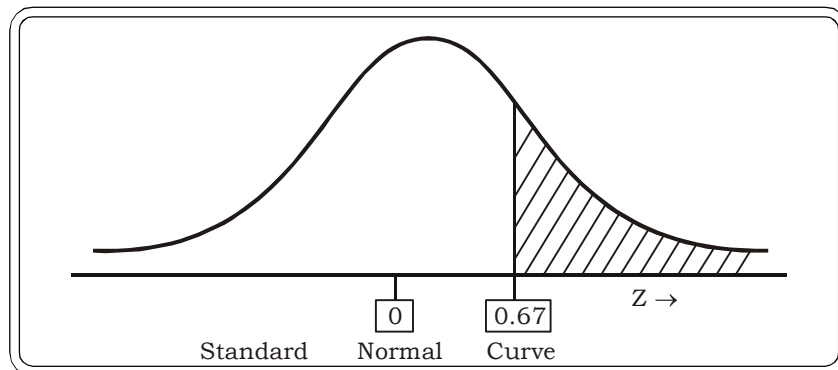
$$= (0.2789) (800)$$

$$\cong 223$$

(ii) $P(X > 160)$

Converting to the Z scale

$$P\left(Z > \frac{160-150}{15}\right) = P(Z > 0.67)$$



$$= 0.5 - P(0 < Z < 0.67)$$

$$= 0.5 - 0.2486$$

$$= 0.2514$$

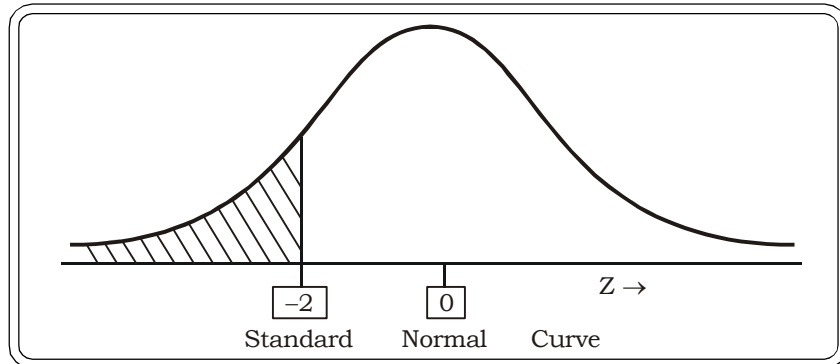
The number of stores with sales more than Rs.160 thousand

$$= (0.2514) (800)$$

$$\cong 200 \text{ stores}$$

(iii) $P(X < 120)$

$$\begin{aligned}
 &= P\left(Z < \frac{120 - 150}{15}\right) = P(Z < -2) = P(Z > 2) \quad (\text{Since the normal curve is symmetric}) \\
 &= 0.5 - P(0 < Z < +2) \\
 &= 0.5 - 0.4772 \\
 &= 0.0228
 \end{aligned}$$



Thus, the number of stores with sales less than Rs.120 thousand

$$\begin{aligned}
 &= (0.0228) (800) \\
 &= 18.24 \\
 &\cong 18
 \end{aligned}$$

Example 5.82: A company manufactures bolts. The specifications are that the bolts should be within 2.99 inches and 3.01 inches in diameter. The process is set to manufacture bolts of 3 inches in diameter. If 5 percent of the bolts are rejected for being out of the upper specification limit, find the standard deviation of the process. Assume that the diameters are normally distributed.

Solution:

Let X – bolt diameter

Then X follows a normal distribution with mean 3 inches.

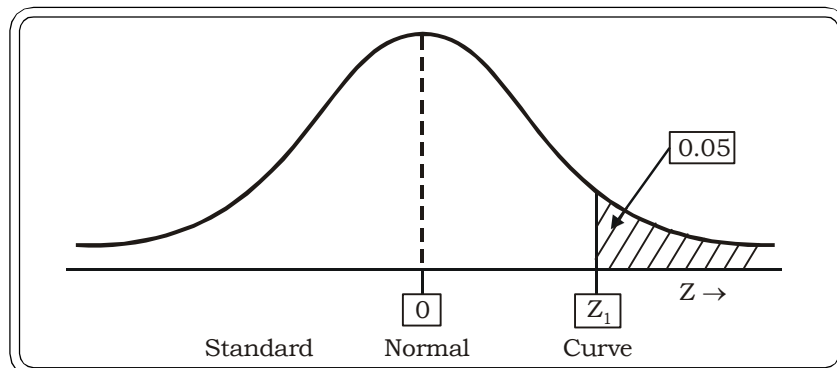
Given 5% of the bolts are out of the upper specification limit

$$P(X > 3.01) = 0.05$$

Converting to the Z-scale:

$$P\left(Z > \frac{3.01 - 3}{\sigma}\right) = 0.05 \quad \dots (1)$$

We have to determine σ



From the standard normal tables, we look for the point above which the area in the curve is 0.05

Or

Alternatively the point Z_1 such that

$$P(0 < Z < Z_1) \\ = 0.5 - .05 = 0.45$$

This point is 1.645

$$\& P(Z > 1.645) = 0.05 \quad \dots (2)$$

Thus comparing (1) & (2)

$$1.645 = \frac{301 - 3}{\sigma}$$

$$\sigma = 0.006$$

Thus standard deviation of the bolts = 0.006 inches

Example 5.83: A large mall has designed a parking lot which has a capacity to park 10000 vehicles on an average with a standard deviation of 2000 cars.

(i) Probability that the no. of cars parked is between 9000 and 11000.

(ii) Probability that the no. of carts parked exceeds 12000.

Solution:

Let X denote the number of vehicles parked.

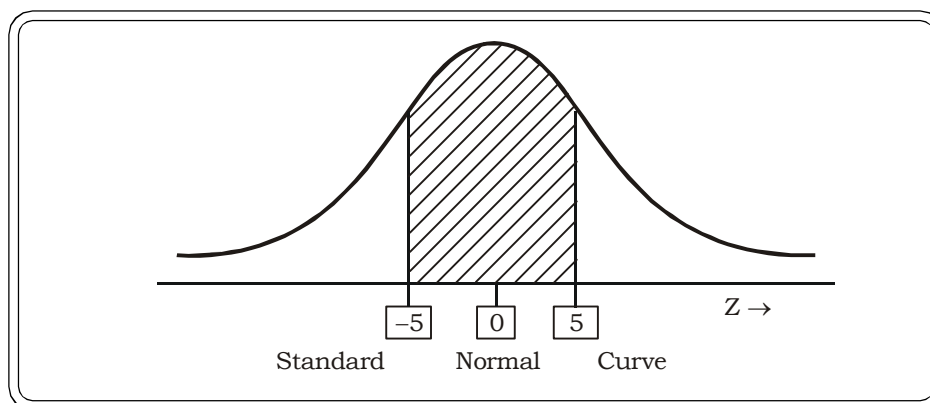
Then, X follows a normal distribution with mean = 10000 cars

And standard deviation = 2000 cars

(i) $P(9000 < X < 11000)$

Converting to the Z scale

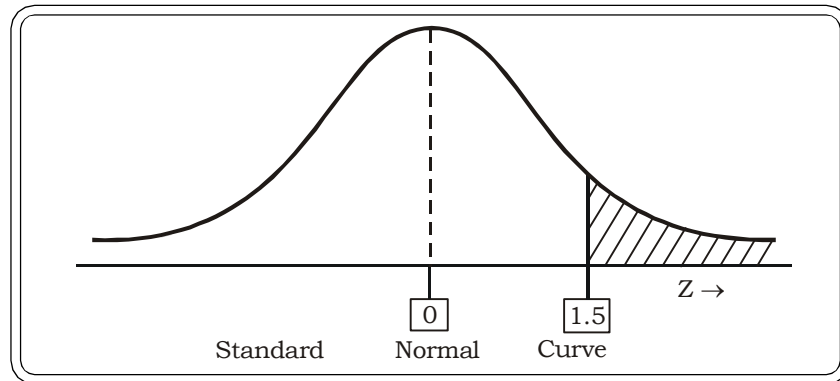
$$= P\left(\frac{9000 - 10000}{2000} < Z < \frac{11000 - 10000}{2000}\right) \\ = P(-0.5 < Z < 0.5)$$



$$\begin{aligned}
 &= 2 P(0 < Z < 0.5) \\
 &= 2 (0.1915) \text{ from tables} \\
 &= 0.383
 \end{aligned}$$

$$(ii) P(X > 12000)$$

$$\begin{aligned}
 &= P\left(Z > \frac{12000 - 9000}{2000}\right) \\
 &= P(Z > 1.5)
 \end{aligned}$$



$$\begin{aligned}
 &= 0.5 - P(0 < Z < 1.5) \\
 &= 0.5 - 0.4332 \\
 &= 0.0668
 \end{aligned}$$

Example 5.84: A company manufacturers detergents in packs of 500gms. The quantity of detergents packed follows a normal distribution with mean 500gms, and standard deviation 2 gms. In a lot of 1000 packets, find the

- (i) No. of packets with weight exceeding 501gms.
- (ii) No. of packets with weight less than 498gms

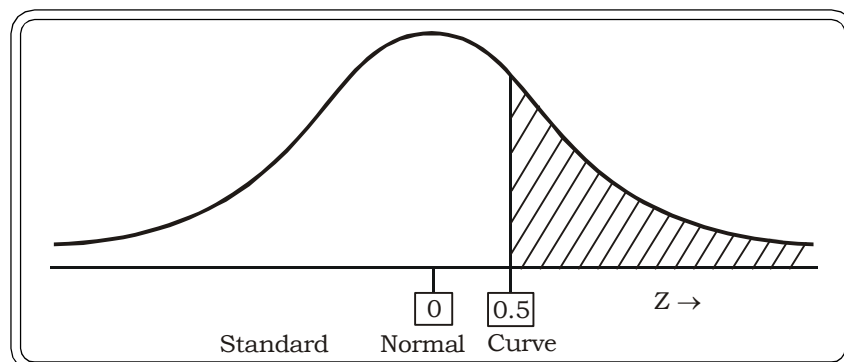
Solution:

Let X – weight of detergent packet

Then X follows a normal distribution with $\mu = 500\text{gms}$ and $\sigma = 2$ gms.

$$(i) P(X > 501)$$

$$\begin{aligned}
 &= P\left(Z > \frac{501 - 500}{2}\right) \\
 &= P(Z > 0.5)
 \end{aligned}$$



$$= 0.5 - P(0 < Z < 0.5)$$

$$= 0.5 - 0.1915$$

$$= 0.3085$$

Thus, in a lot of 1000 packets, the no. of packets exceeding 501 gms

$$= (0.3085) (1000)$$

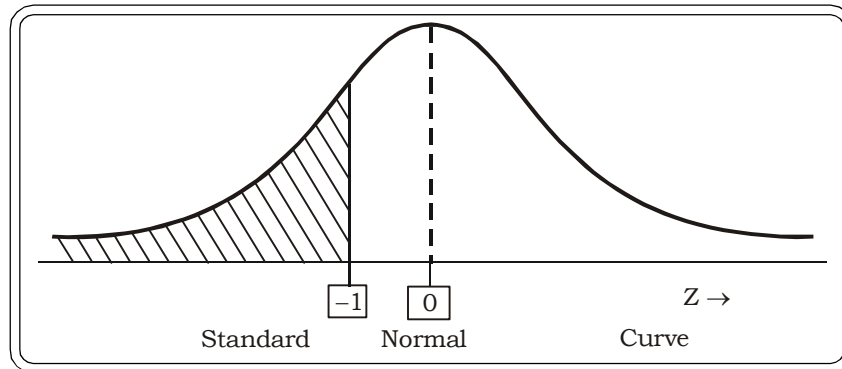
$$= 308.5$$

$$\cong 309 \text{ packets}$$

(ii) $P(X < 498)$

$$= P\left(Z < \frac{498 - 500}{2}\right)$$

$$= P(Z < -1)$$



$$= P(Z > 1), \because \text{the curve is symmetric}$$

$$= 0.5 - P(0 < Z < 1)$$

$$= 0.5 - 0.3413$$

$$= 0.1587$$

Thus, in a lot of 1000, number of packets with weight less than 498gms is

$$= (0.1587) (1000)$$

$$= 158.7$$

$$\cong 159 \text{ packets}$$

5.9.3 Normal as an approximation to Binomial Distribution

For large values of n , the binomial distribution can be approximated by a normal distribution, provided p is not close to 0 or 1. The mean of the approximate normal distribution is np and the standard deviation is \sqrt{npq} .

The question arises as to how a discrete distribution can be approximated by a continuous distribution. The following illustration explains this point.

Consider a binomial distribution with $n = 15$ and $p = 0.4$

$$P(X = 6) = 0.217 \text{ (from binomial tables)}$$

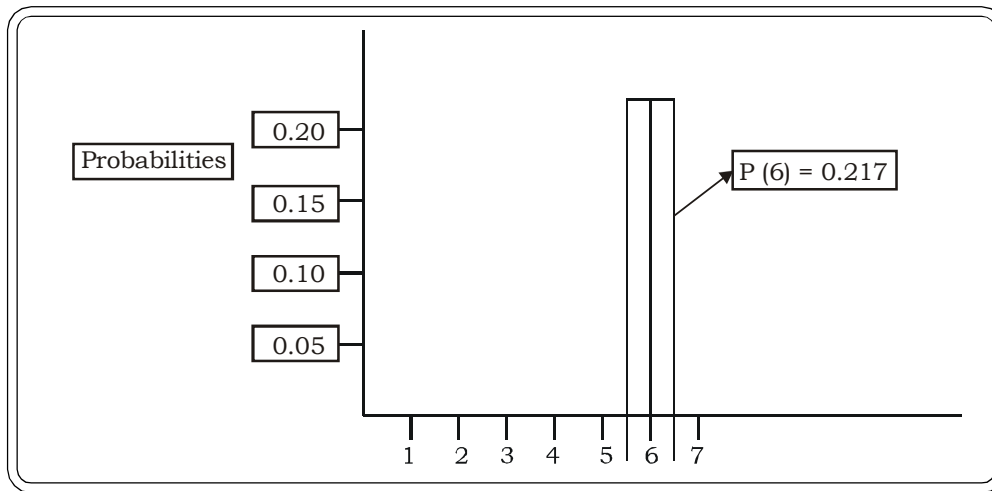


Figure 5.17

Graph of Binomial Probability $P(6) = 0.217$

From the graph consider the histogram centered at $X = 6$ and having a base length of 1 unit from 5.5 to 6.5. The area of this histogram = $0.217 \times 1 = 0.217$

Thus, the binomial probability of $P(6)$ is equal to the area of the rectangle in the figure.

Suppose we approximate this probability by a normal distribution

$$\mu = np = (15)(0.4) = 6$$

$$\sigma = \sqrt{npq} = 1.897$$

$$\begin{aligned} P(5.5 < X < 6.5) &= P\left(\frac{5.5-6}{1.897} < Z < \frac{6.5-6}{1.897}\right) \\ &= P(-0.26 < Z < 0.26) \end{aligned}$$

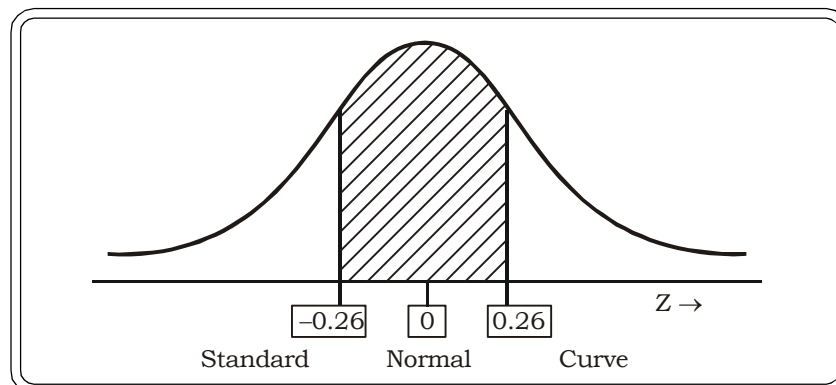


Figure 5.18

Standard Normal Probability Curve for $P(5.5 < X < 6.5)$

$$= 2 P(0 < Z < 0.26)$$

$$= 2 (0.126)$$

$$= 0.252$$

which is close to 0.217, the binomial probability obtained earlier.

Thus, a normal distribution gives a close approximation to the binomial distribution, when n is large and p is not close to 0 or 1.

Example 5.85: A fast food owner calculates that about 40% of his customers order soft drink with burgers. If 500 customers visit the fast food joint, what is the probability that more than 190 will order soft drinks with burgers?

Solution:

Let X – number of customers who order soft drink with burgers
 p – probability that customer orders soft drinks with burgers
 $= 0.40$
 $n = 500$

We have to find $P(X > 190)$

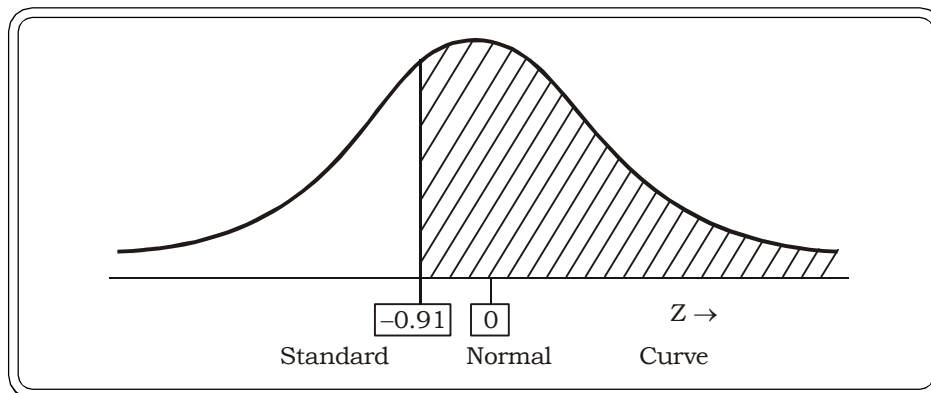
Since n is large we have to use the normal approximation to the binomial distribution

$$\mu = np = 500(0.40) = 200, \sigma^2 = npq = 500(0.40)(0.60) = 120, \sigma = 10.95$$

$$P(X > 190) = P\left(Z > \frac{190 - 200}{10.95}\right)$$

Converting to the standard Z - scale

$$= P(Z > -0.91)$$



$$= 0.5 + P(0 < Z < 0.91)$$

$$= 0.5 + 0.3186 = 0.8186$$

Thus the approximate probability that more than 190 customers will order soft drinks with their burgers is 0.8186

Example 5.86: A machine produces bolts of a certain type. In a sample of 100 bolts, if the number of defectives is less than 12, the entire production lot is accepted. Find the probability that the lot is accepted when the machine produces 20% defective bolts.

Solution:

Let X – number of defective bolts
 $p = 0.20$ = probability of a defective bolt.
 $n = 100$

The lot will be accepted if the no. of defectives is < 12

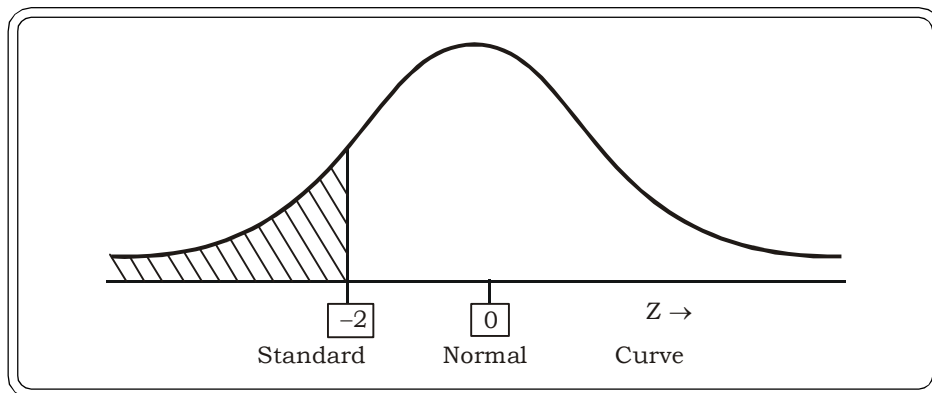
Approximating by a normal distribution, since the sample size is large.

$$\mu = np = (100)(0.20) = 20$$

$$\sigma = \sqrt{npq} = 4$$

$$P(X < 12) = P\left(Z < \frac{12 - 20}{4}\right)$$

$$= P(Z < -2)$$



$$= 0.5 - P(0 < Z < 2)$$

$$= 0.5 - 0.4772 = 0.0228$$

Thus, probability that the production will be accepted is 0.0228

Example 5.87: On an average, 5% of all televisions coming in for final inspection in a factory fail to pass. Use the normal approximation to the binomial to find the probability that between 7 to 12 of the next batch of 200 TV's coming in for final inspection will fail the test.

Solution:

Let X – the number of TV's that fail to pass the final inspection

p – probability of failing the final inspection

$$= 0.05$$

$$n = 200$$

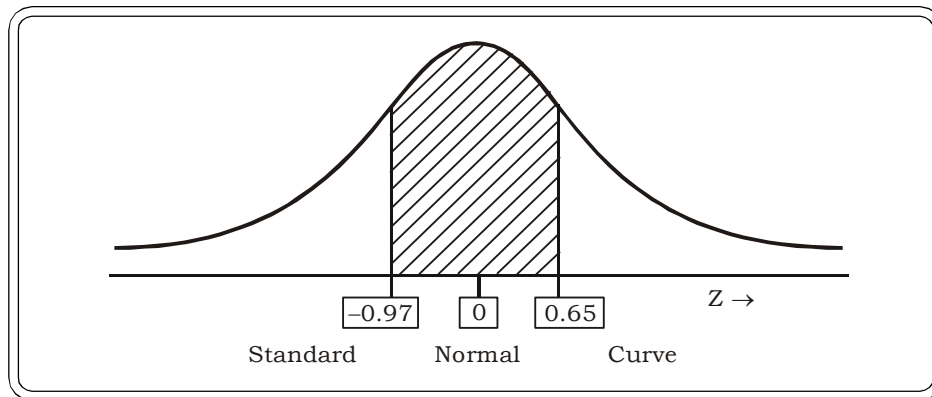
We find the mean and variance of the normal distribution to approximate the binomial distribution

$$\mu = np = (200)(0.05) = 10$$

$$\sigma = \sqrt{npq} = 3.08$$

$$P(7 < X < 12) = P\left(\frac{7 - 10}{3.08} < Z < \frac{12 - 10}{3.08}\right)$$

$$= P(-0.97 < Z < + 0.65)$$



$$= P(0 < Z < 0.65) + P(0 < Z < 0.97)$$

$$= (0.2400) + (0.3340) = 0.574$$

$$\text{Thus } P(7 < X < 8) = 0.574$$

5.10 CASELET

End of Euro Kids

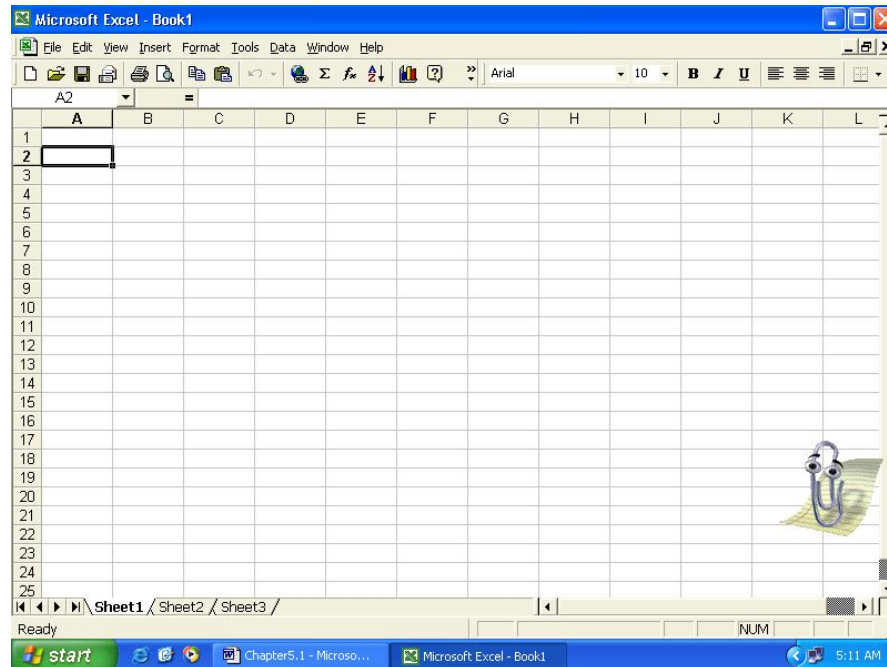
The fast declining, birth rate is becoming a cause of concern for the EU countries. During communist lines, a clinic in a German city used to deliver on average 2500 babies a year with a standard deviation of 500 babies. Currently the average has fallen to 800 with a standard deviation of 100. In the tumultuous years after World War II, Europe's birth rate went up, peaking in the mid 60's. Since then the trend has been on a downswing. According to statistics Germany has the lowest birthrate in the EU with 8.5 births per 1000 inhabitants. The average number of babies born in Germany was 6,50,000 with standard deviation of 50,000 it is the lowest since 1945. Other nations at the bottom of the table include Spain (10.6 per 1000 inhabitants), Italy (9.7) and Poland (9.3). In Europe 2.1 is considered to be the population replacement ratio. This is the average number of children per women. However major European countries like Ireland (1.99), U.K. (1.74), Germany (1.37), Italy (1.33), Greece (1.29) has a replacement ratio much lower than the average. The alarming statistics have ensured that European's birth rate has become a matter of urgent political concern. In Germany, a tabloid recently predicted that Germans would die out entirely by 2300. (*Adapted from Hindustan Times, June 7, 2006*)

Use techniques discussed in this chapter to analyze this situation. Also give suggestions to arrest this problem. State any assumptions that need to be made.

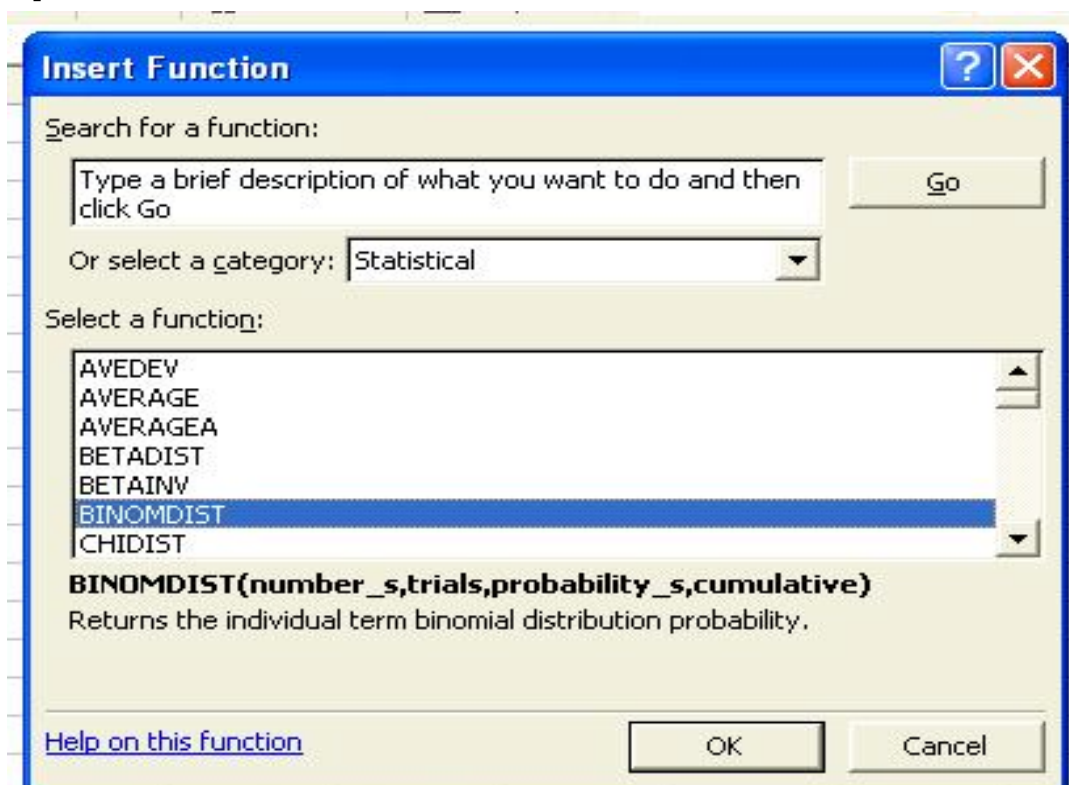
5.11 EXCEL GUIDE

Using Excel to calculate probabilities of Binomial Distribution

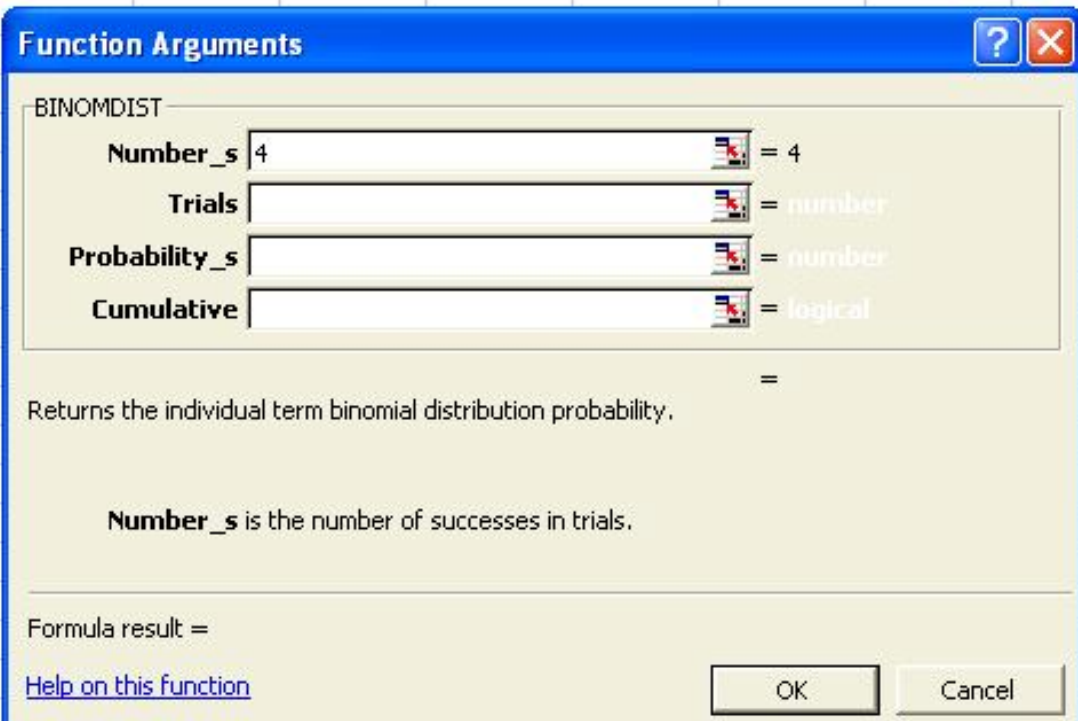
Step 1: Select a cell where you want the result



Step 2: Go to INSERT FUNCTION OPTION and select the category STATISTICAL in SELECT A FUNCTION option. Select BINOMDIST. Click OK.



Step 3: In function argument, first ENTER the number of successes. Here we enter 4. Then go to TRIALS.




The image shows the "Function Arguments" dialog box for the BINOMDIST function. The dialog has a blue title bar with a question mark and a close button. The main area is light yellow and contains the following fields:

- Number_s**: 4
- Trials**: (empty)
- Probability_s**: (empty)
- Cumulative**: (empty)

Each field has a small icon to its right. Below the fields, there is a description: "Returns the individual term binomial distribution probability." and a note: "Number_s is the number of successes in trials." At the bottom, there is a "Formula result =" field, a "Help on this function" link, and "OK" and "Cancel" buttons.

Step 4: Enter the number of trials - 6 in this case.



The image shows the "Function Arguments" dialog box for the BINOMDIST function, similar to the previous one, but with the "Trials" field now containing the value 6. The other fields remain empty. The description and note are the same as in the previous dialog box. The "Formula result =" field, "Help on this function" link, and "OK" and "Cancel" buttons are also present.

Step 5: Enter the probability of success in each trial : 0.4 in this case.

Function Arguments

BINOMDIST

Number_s 4 = 4

Trials 6 = 6

Probability_s 0.4 = 0.4

Cumulative = logical

=

Returns the individual term binomial distribution probability.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability mass function, use FALSE.

Formula result =

[Help on this function](#)

Step 6: In cumulative type FALSE for individual probability value & TRUE for the cumulative probabilities up to that number. In this case we have entered FALSE, so that we get the binomial probability of p (4). Finally clicks OK.

Function Arguments

BINOMDIST

Number_s 4 = 4

Trials 6 = 6

Probability_s 0.4 = 0.4

Cumulative false = FALSE

= 0.13824

Returns the individual term binomial distribution probability.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability mass function, use FALSE.

Formula result = 0.13824

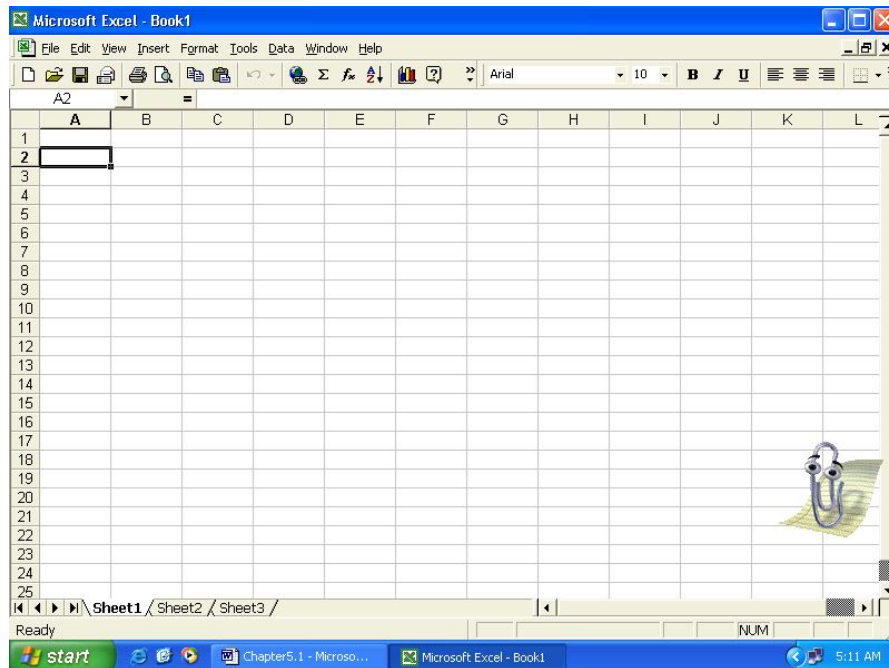
[Help on this function](#)

Step 7: The binomial probability $p(4) = 0.13824$ is displayed in cell A2.

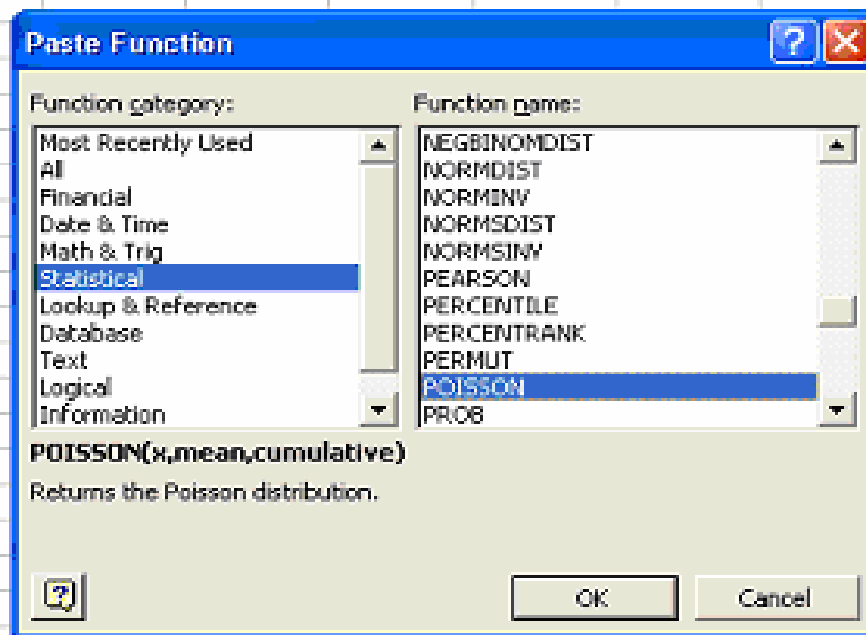
	A	B	C
1			
2	0.13824		
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			

Using Excel to calculate probabilities of Poisson Distribution

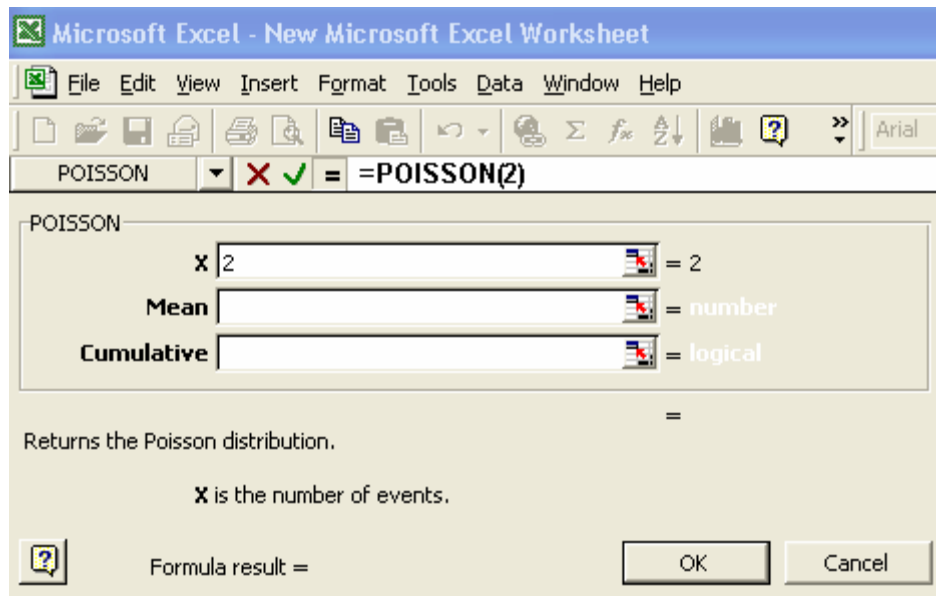
Step 1: Select a cell where you want the result



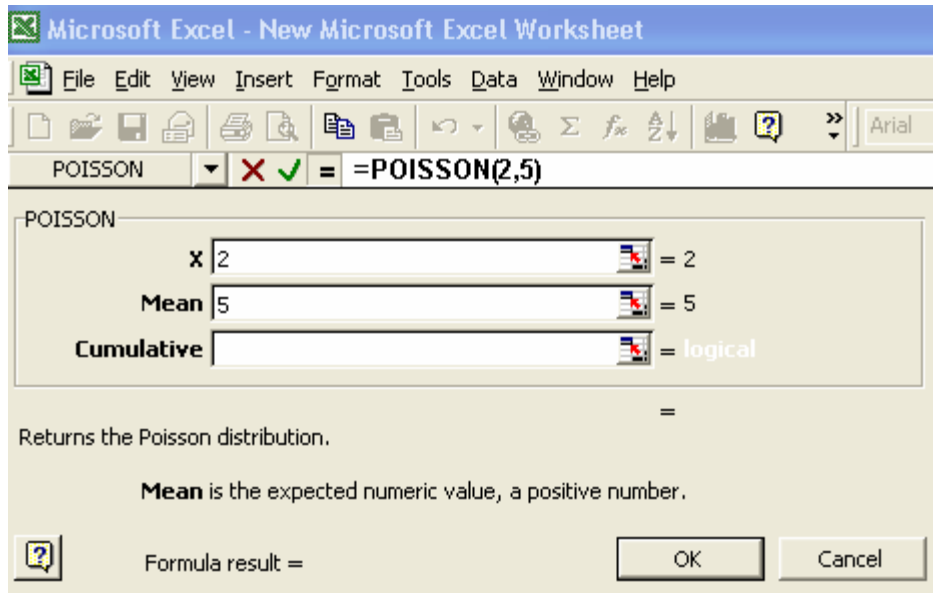
Step 2: Go to INSERT FUNCTION OPTION and select the category STATISTICAL in SELECT A FUNCTION option. Select POISSON. Click OK



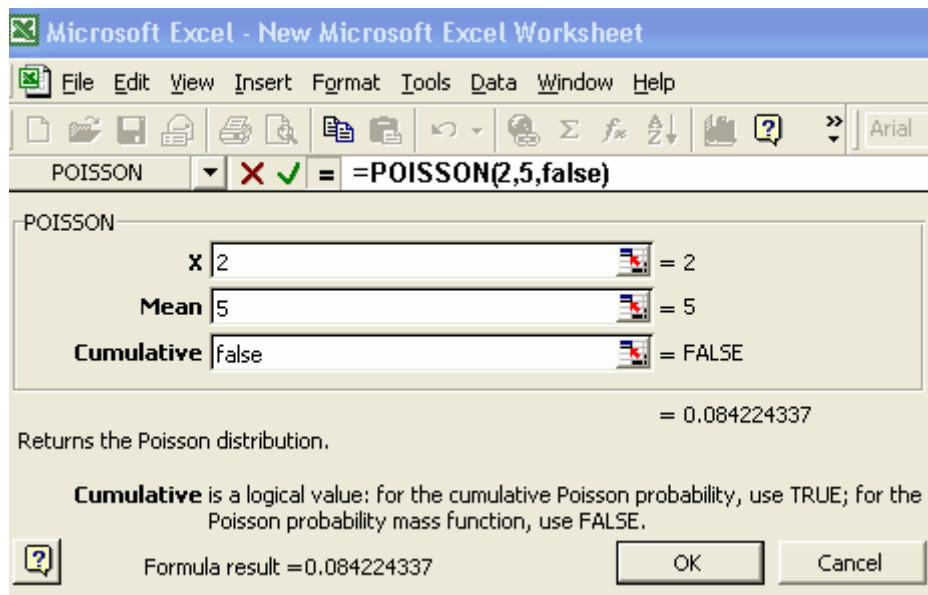
Step 3: Enter the value of x – the number of events. Here we enter x = 2



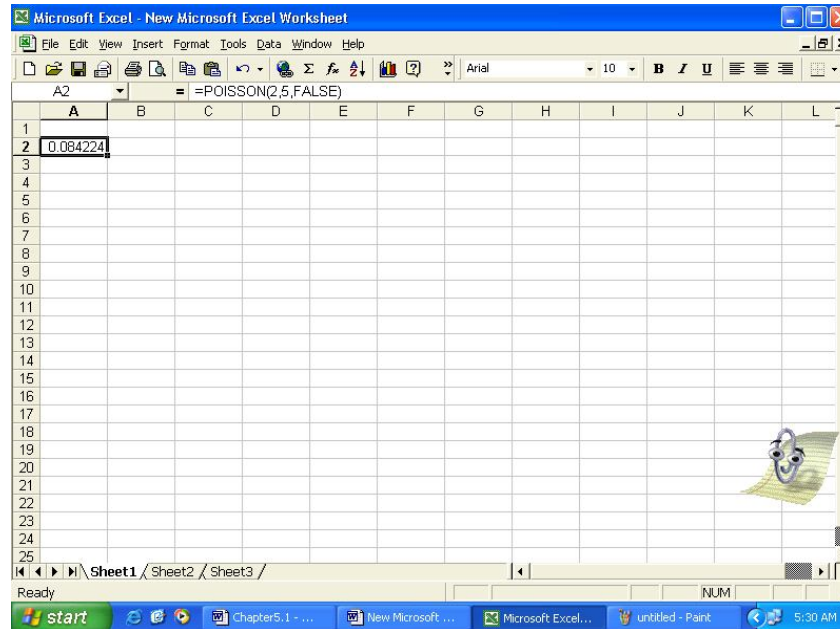
Step 4: The mean of the Poisson distribution is now entered as 5.



Step 5: In cumulative type FALSE for individual probability value & TRUE for the cumulative probabilities up to that number. In this case we have entered FALSE, so that we get the Poisson probability of p (2). Click OK.

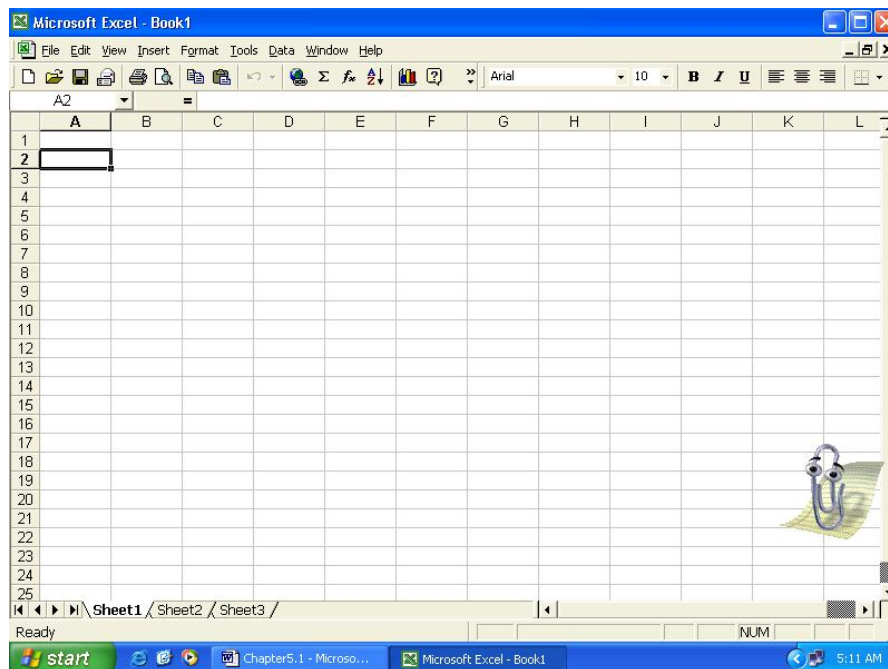


Step 6: The Poisson probability $p(2) = 0.0842$ is displayed in cell A2

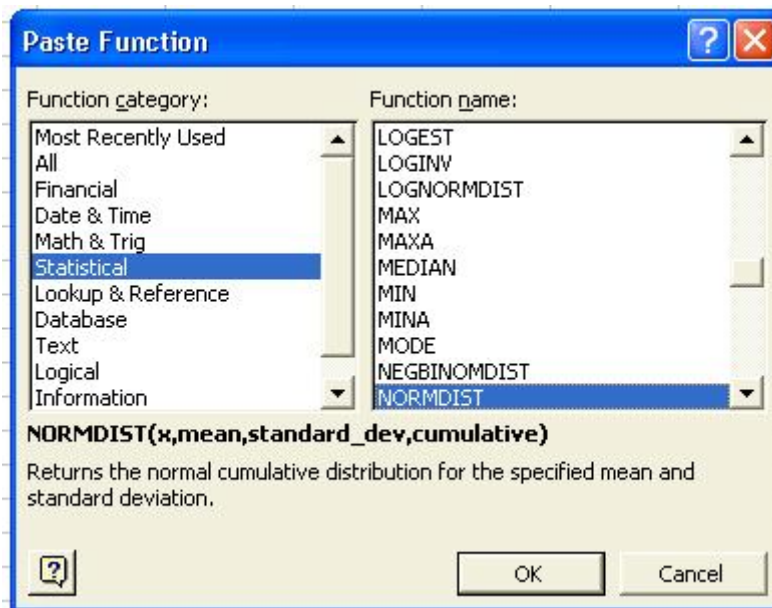


Using Excel to calculate probabilities of Normal Distribution

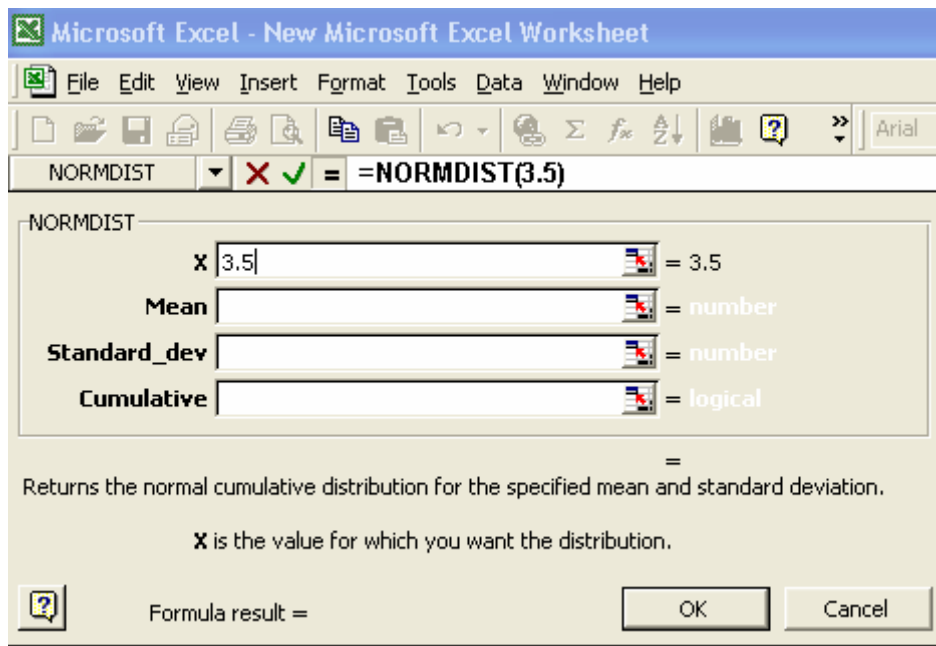
Step 1: Select a cell where you want the result



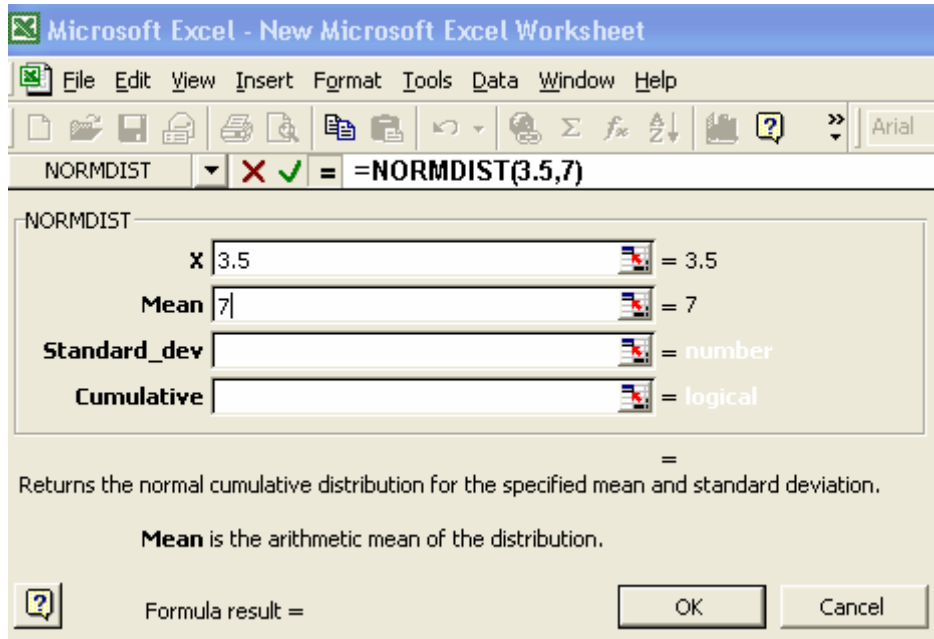
Step 2: Step 2: Go to INSERT FUNCTION OPTION and select the category STATISTICAL in SELECT A FUNCTION option. Select NORMDIST. Click OK



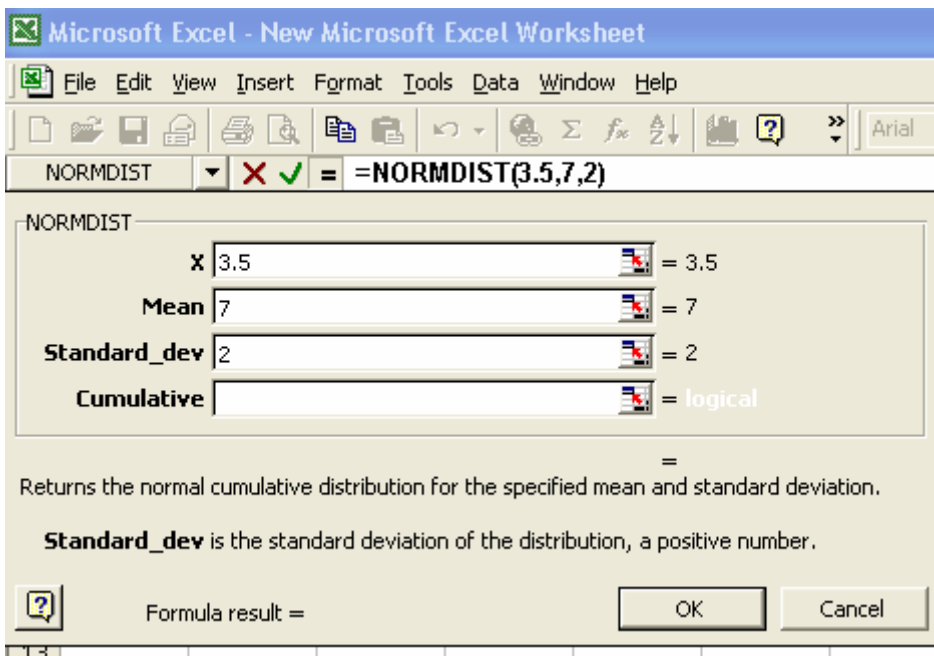
Step 3: Enter x as 3.5



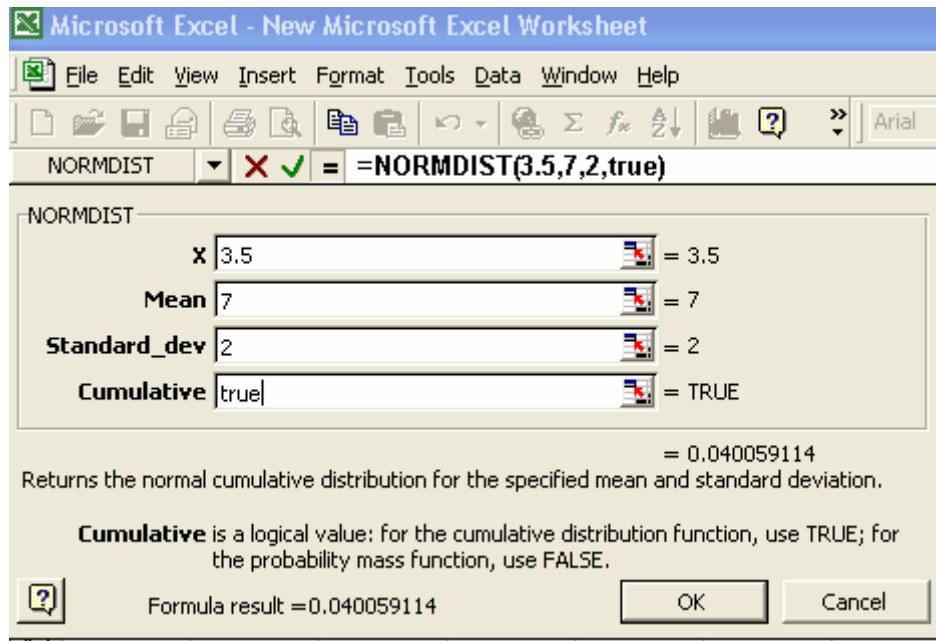
Step 4: Enter the mean of the normal distribution. Here mean = 7.



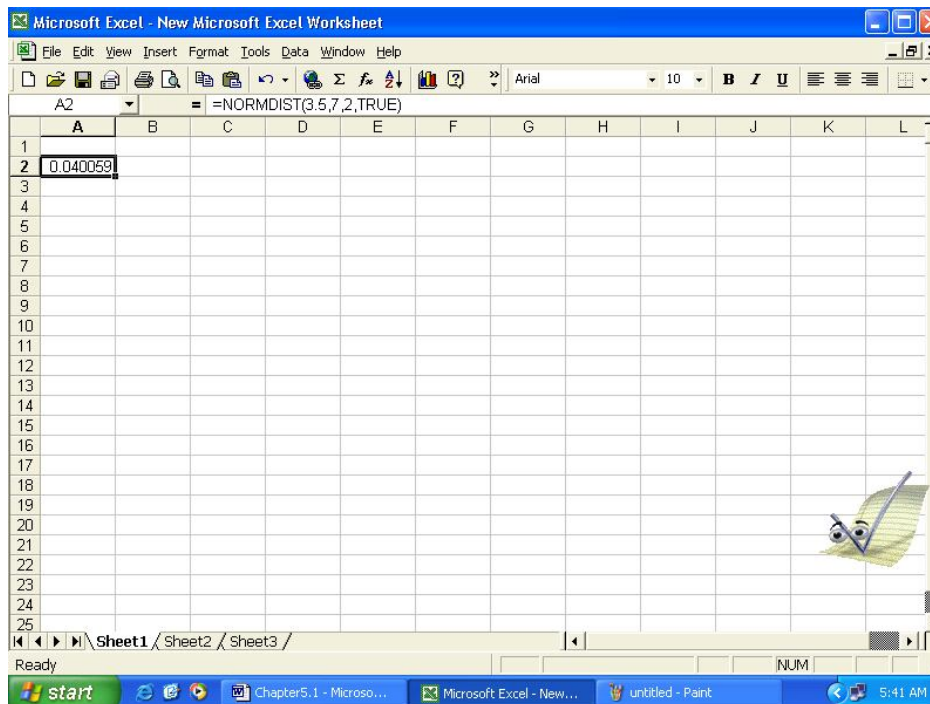
Step 5: Enter the standard deviation of the normal distribution. Here S.D. = 2



Step 6: Enter TRUE for $p(x < 3.5)$ and FALSE for $p(x = 3.5)$. click OK



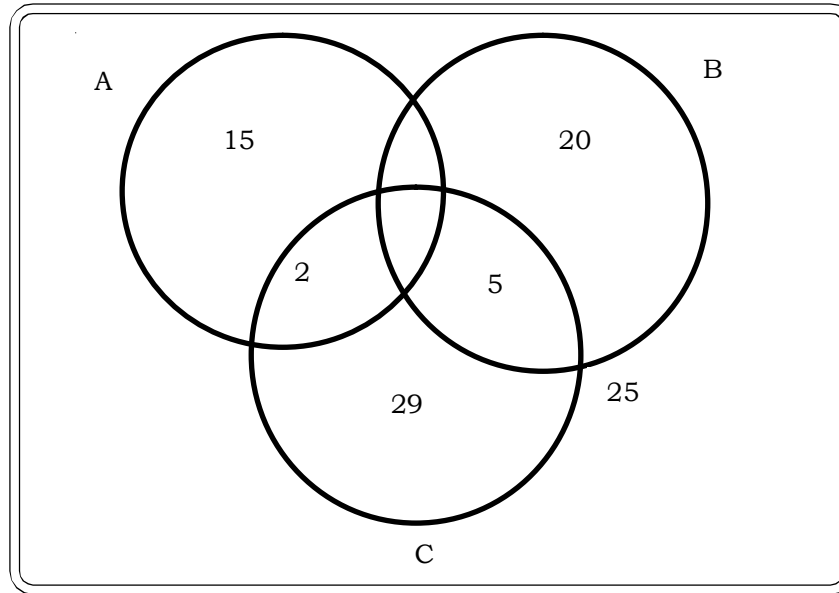
Step 7: $p(x < 3.5) = 0.40059$ is displayed in cell A2.



Step 8: For probability mass function, we type FALSE in step 6.

5.12 EXERCISES

- 5.1 If $P(A) = 0.5$, $P(B) = 0.4$ and $P(A \text{ and } B) = 0.28$. find $P(A \text{ or } B)$. Also show this in a Venn diagram.
- 5.2 From the following Venn diagram. Calculate $P(A)$, $P(B)$, $P(A \text{ or } B)$, $P(C)$.



Given $P(ABC) = 2$, $P(AB) = 4$

- 5.3 Give the classical and the Empirical definitions of probability.
- 5.4 Explain the axiomatic approach of probability.
- 5.5 Define the following terms:
- (i) Sample Space
 - (ii) Trial or Experiment
 - (iii) Event
- 5.6 What do you mean by mutually exclusive events? Give an example of 2 mutually exclusive events.
- 5.7 Define independent events with examples.
- 5.8 Can mutually exclusive events also be independent? Explain with examples.
- 5.9 Describe briefly the various schools of thought on probability. How does the concept of probability help decision – maker to improve his decisions?
(MBA, HPU, 1998; MBA, DU, 1999)
- 5.10 Define independent and mutually exclusive events. Can two events be mutually exclusive and independent simultaneously? Support your answer with an example.
(MBA, Sukhadia Univ., MBA, DU, 1999)

- 5.11 State the Baye's Theorem.
- 5.12 A problem in statistics is given to the three students A, B and C whose chances of solving it are $1/2$., $3/4$., and $1/4$ respectively. What is the probability that the problem will be solved if all of them try independently? **(Madurai Kamraj Univ., B.Sc., 1986; DU, B.A., 1991)**
- 5.13 Three groups of children contain respectively 3 girls and 1 boy, 2 girls and 2 boys, and 1 girl and 3 boys. One child is selected at random from each group. Show that the chance that the three selected consists of 1 girl and 2 boys is $13/32$?
(Madurai Univ., B.Sc, 1998; Nagpur Univ., B.Sc., 1991)
- 5.14 In 1989 there were three candidates for the position of principal – Mr. Chatterji, Mr. Ayangar and Dr. Singh – whose chances of getting the appointment are in the proportion 4:2:3 respectively. The probability that Mr. Chatterji if selected would introduce co – education in the college is 0.3. The probabilities of Mr. Ayangar and Dr. Singh doing the same are respectively 0.5 and 0.8. What is the probability that there was co – education in the college in 1990? **(DU, B.Sc (Stat. Hons.), 1992; Gorakhpur Univ. B.Sc., 1992)**
- 5.15 The contents of urns I, II and III are as follows:
- 1 white, 2 black and 3 red balls,
 - 2 white, 1 black and 1 red balls, and
 - 4 white, 5 black and 3 red balls
- One urn is chosen at random and two balls drawn. They happen to be white and red. What is the probability that they come from urns I, II and III? **(DU, B.Sc. (stat. Hons.), 1998)**
- 5.16 Two computers A and B are to be marked. A salesman who is assigned the job of finding customers for them has 60% and 40% chances respectively of succeeding in case of computers A and B. The 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?
(MBA, IGNOU, 2002; MBA, DU, 2002)
- 5.17 The Human Resource department of a company has records, which show the following analysis of its engineers.

Age	Bachelor's degree only	Mater'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

- (i) The probability he has only a bachelor's degree
- (ii) The probability he has a master's degree, given that he is over 40
- (iii) The probability he is under 30, given that he has only a bachelor's degree

5.18 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 was obtained are summarized in the following cross tabulation.

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
Total	8	10	11	8	3	40

(i) Develop a joint probability table

(ii) What is the probability of one of the participants studied had a satisfaction score in the 80's?

(iii) What is the probability of a satisfaction score in the 80's given the study participant was a doctor?

(iv) What is the probability of one of the participants studied was a lawyer?

(v) What is the probability of one of the participants was a lawyer and received a score under 50?

(vi) What is the probability of a satisfaction score under 50 given a person is a lawyer?

(vii) What is the probability of satisfaction score of 70 or higher? **(MBA, DU, 2003)**

5.19 Project Vijay, NCSO, INDIA sums its operations on 10 computers, which may need repairs from time to time during the day. Three of these computers are old, each having a probability of $1/11$ of needing repair during the day and seven are new, having corresponding probability of $1/21$. Assuming that no computers needs repair twice on the same day, determine the probabilities that on a particular day.

(i) Just 2 old and no new computers need repair.

(ii) If just 2 computers need repair, they are of same type. **(MBA, IGNOU, Dec. 2000)**

5.20 Explain Bernoulli trials and the conditions that are a pre requisite for Bernoulli trials.

5.21 Derive the binomial probabilities from that of Bernoulli trials. State the assumptions made.

5.22 Difference between discrete and continuous random variable with examples.

5.23 The mean of a Binomial distribution is 4 and its variance is 2. Find n, p and q.

5.24 The mean of a binomial distribution is 5 and the variance is 10. Is this statement is true? Explain why or why not?

5.25 State the important properties of a normal distribution.

5.26 Under what conditions can a binomial distribution be approximated by a normal distribution?

5.27 X follows a normal distribution with mean 10 and variance 4. Find the following probabilities

- (i) $P(0 \leq X \leq 11)$
- (ii) $P(X \geq 10.5)$
- (iii) $P(X \leq 12)$
- (iv) $P(8 < X < 12)$
- (v) $P(X \geq 9.6)$
- (vi) $P(8.2 < X < 11.8)$

5.28 Let Z can be a standard normal variable. Find the value of c if

- (i) $P(0 < Z < C) = 0.1844$
- (ii) $P(Z \geq C) = 0.7357$
- (iii) $P(C < Z < 2) = 0.64$
- (iv) $P(Z \leq C) = 0.8729$

5.29 The number of orders that a bakery receives in a month for birthday cakes is approximately normally distributed with mean number of orders equal to 500 and standard deviation of 30 orders. Find the probability that the number of orders received in a month.

- (i) Exceeds 440
- (ii) Is less than 450
- (iii) Between 440 and 470

5.30 The incidence of occupational disease in a cement manufacturing industry is such that a worker has a 20% chance of suffering from it. Find the probability that 8 out of 10 workers will have the disease.

5.31 Suppose that a manufactured product has 2 % chance of being defective. In a lot of 12 products find the probability that less than 2 products are defective.

5.32 One hundred car radio sets are inspected as they come off the production line and number of defects per set is recorded below:

No. of defects	0	1	2	3	4
No. of sets	79	18	2	1	0

Fit a Poisson distribution to the above data.

(MBA, DU, 1999)

5.33 A manufacturer who produces medicine bottles, finds that 0.1% of the bottles are defective. The bottles are packed in boxes containing 500 bottles. A drug manufacturer buys 100 boxes from the producer of bottles. Using Poisson distribution, find how many boxes contain:

- (i) No defectives
- (ii) At least two defectives

(MBA, DU, 2001)

- 5.34 Suppose that half of the population of town are consumers of rice. One hundred investigators are appointed to find out its truth. Each investigator interviewed 10 individuals. How many investigators do you expect to report three or less of the people interviewed are consumers of rice?
(MBA, Bharathidasan Univ., Nov. 2001)
- 5.35 Eight coins are thrown simultaneously. Using binomial distribution, show that the probability of obtaining at least 6 heads is 0.1445.
(MBA, DU, 2003)
- 5.36 The probability that a door – to – door salesman makes a sale is 0.25. In a 50 house calls, find the expected number of houses in which he is expected to make a sale and the variance of the number of houses where he makes a sale.
- 5.37 The local authorities in a certain city install 10, 000 electric lamps in the streets of the city. If these lamps have an average life of 1000 burning hours with a standard deviation of 200 hours, assuming normality, what number of lamps might be expected to fail in the first 800 burning hours?
(MBA, IGNOU, Dec, 2001)
- 5.38 A TV channel has indicated that 80% of all families watch their program 'Musically yours' telecast every Saturday at 9:00 p.m. In a sample of 15 households what is the probability that 10 households watch this program.
- 5.39 The salaries of MBA graduates who pass out of a certain university follows a normal distribution with mean salaries Rs.50, 000 per month with a standard deviation of Rs. 20,000. Find in a batch of 200 how many students will have salaries exceeding Rs.40, 000 per month.
- 5.40 The probability of errors in credit card statements of a particular bank is 0.20. In a survey of 1000 account statements, what is the probability that the no. of statements with errors exceeds 300. (use normal approximation of the binomial distribution)



6

Sampling and Sampling Distributions



Structure

- 6.1 Introduction
- 6.2 Parameter and Statistic
- 6.3 Sampling: Meaning, Steps and Types of Sampling
 - 6.3.1 Probability Sampling Methods
 - 6.3.1.1 Simple Random Sampling
 - 6.3.1.2 Stratified Random Sampling
 - 6.3.1.3 Systematic Sampling
 - 6.3.1.4 Cluster Sampling
 - 6.3.1.5 Multistage Sampling
 - 6.3.2 Non - Probability Sampling Methods
 - 6.3.2.1 Judgement Sampling
 - 6.3.2.2 Convenience Sampling
 - 6.3.2.3 Quota Sampling
 - 6.3.3 Sampling and Non-Sampling Errors
- 6.4 Sampling Distributions
 - 6.4.1 The Central Limit Theorem
 - 6.4.2 Sampling Distribution of the Mean
 - 6.4.3 Sampling Distribution of the Proportion
 - 6.4.4 Student's t-Statistic and its Distribution
 - 6.4.5 The Chi-Square Statistic and its Distribution
 - 6.4.6 The F-Statistic and its Distribution
- 6.5 Exercises

6.1 INTRODUCTION

Inferential or inductive statistics is primarily concerned with making conclusions about a certain population or populations. Since in most cases it is not possible to do a 100% inspection or complete census, the logical thing to do is to take a sample from the population. This sample is then studied and the results of the study are generalized to include the entire population. Apart from a complete enumeration being infeasible, studying a sample also has other advantages. Studying a sample is faster and the cost involved is also significantly less as compared to a census. Also, the quality of data collected would leave much to be desired if an entire population is being studied. Thus, the entire inferential theory is based on how to make statistically valid conclusions from a sample study.

6.2 PARAMETER AND STATISTIC

All conceivable units of the group being studied are referred to as the population or universe. And a sample is a sub group from the universe under study. For example suppose a machine produces a lot of 1lakh bulbs. The length of life of these bulbs needs to be tested. Since testing the entire lot is practically not possible, we will take recourse to sampling. For example we may select a random sample of 1000 bulbs to decide about the entire lot. So the lot of 1lakh bulbs is the population and the 1000 bulbs selected for testing is the sample.

Now the next question arises as to how and what to test in these bulbs. Since our interest lies in knowing the length of life, the mean life of the bulbs in the sample may provide us with useful information about the mean life of the entire lot. This brings us to the concept of a parameter and a statistic.

A parameter is defined as a quantitative measure that describes a certain characteristic of a population. In our example the ***mean life of the bulbs of the entire lot*** is a **parameter**.

A statistic on the other hand is a quantitative measure, which describes a characteristic of the sample. In this example the ***mean life of the 1000 bulbs in the sample*** is a **statistic**.

As another example consider a poll where people were asked to say “yes” or “no” to the question. “Will India Cricket win the World Cup in 2006?” The poll was aimed to capture the general opinion of Indians about India’s ability to win the cricket world cup. The population or universe is all Indians. We now take a sample of say 1,00,000 people from across the country and out of these 70,000 say “yes” to this question. So the sample proportion is 0.70. We can say that 70% of the sample respondents are confident about India’s ability to win the world cup. The parameter here is the *proportion* of Indians who believe India will win the world cup. The statistic is “*the sample proportion*” i.e. 70% of Indians in the sample who have faith in India’s ability to win.

A point to be noted here is that the value of the statistic will vary from sample to sample. A statistic is thus a random variable. A parameter on the other hand, is a constant peculiar to a given population. We will come back to the implication of this in our section on sampling distributions.

6.3 SAMPLING: MEANING, STEPS AND TYPES OF SAMPLING

In the previous section we discussed concepts like population, sample, parameter and statistic. Sampling may be defined as selecting a part of an aggregate to represent the whole. In other words, it is the selection of a sample to study the universe.

Step in sampling

The sampling process generally consists of the following steps:

- (i) Defining the population
 - (ii) Identification of the sampling frame
 - (iii) Defining the sampling unit
 - (iv) Selecting the sampling method
 - (v) Determining the sample size
 - (vi) Actual sample selection.
- (i) Defining the population: At the beginning of the survey, the population of interest must be clearly defined in terms of its elements.
 - (ii) Identification of the sampling frame: The sampling frame is defined as the list of all elements of the population under study. This frame is used to draw the sample from the population. A sampling frame could be a complete list of households of a given area, a telephone directory or a list of all retailers of a company.
 - (iii) Defining the sampling unit: The sampling unit consists of individual units in the population from which the sample is to be drawn. For example, in a sampling frame consisting of retailers, each retailer is a sampling unit.
 - (iv) Selecting the sampling method: There are various methods of selecting a sample each with their own advantages and disadvantages. These methods are described in section 6.3.1 and 6.3.2.
 - (v) Determining the sample size: Sample size determination depends on a number of factors like.
 - The type of study being conducted. For exploratory research, a small sample size is sufficient, whereas for descriptive research, a large sample may be suitable.
 - Sample size also depends upon the available resources.
 - Sample size is also dependent on the amount of accuracy (or level of confidence) desired in the results.

Types of Sampling

Broadly, there are two methods to select samples from a population viz:

1. Probability Sampling Methods.
2. Non-Probability Sampling Methods.

These are further classified as shown in the following figure:

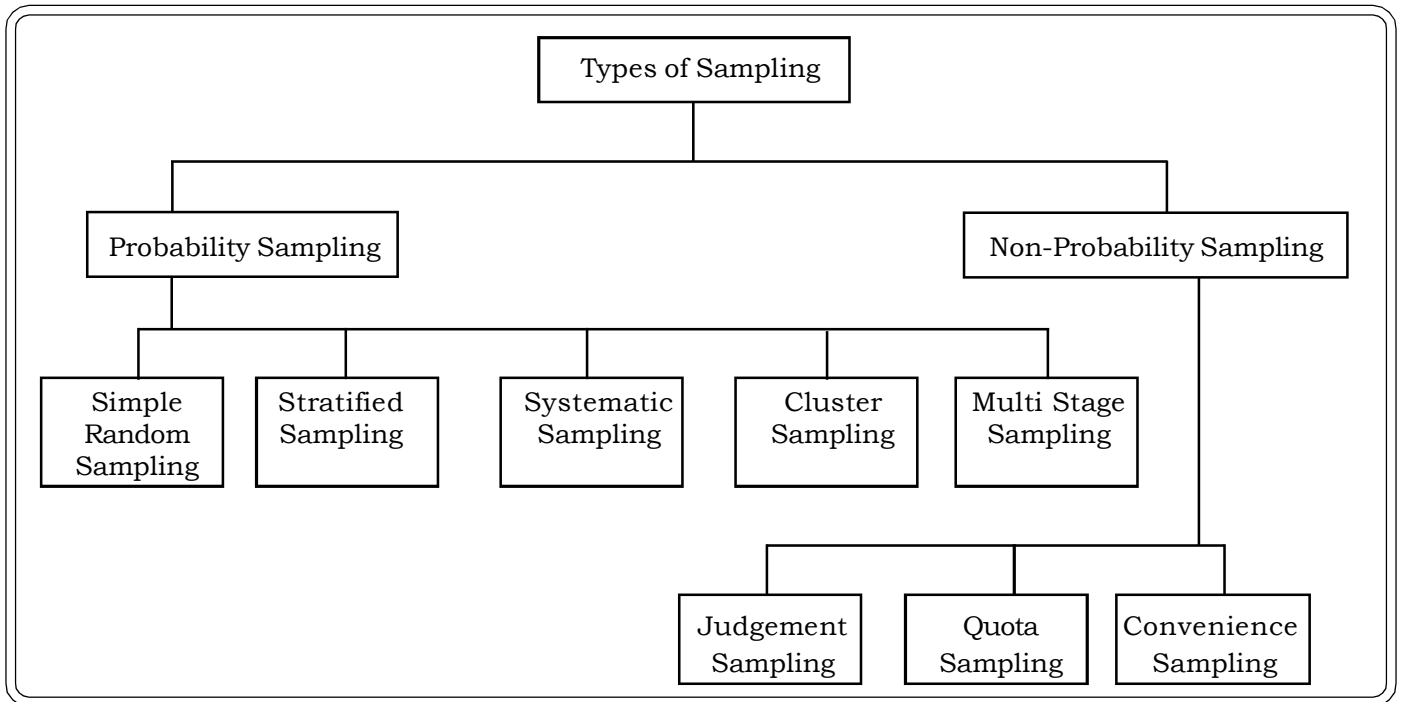


Figure 6.1

Types of Sampling

A brief discussion of these methods now follows:

6.3.1 Probability Sampling Methods

In these methods, each unit in the population has a certain pre assigned probability of being included in the sample. It is a scientific method of selecting a sample.

Some of the probability sampling methods are

- (i) Simple Random Sampling
- (ii) Stratified Random Sampling
- (iii) Systematic Sampling
- (iv) Cluster Sampling
- (v) Multistage Sampling

6.3.1.1 Simple Random Sampling

This is one of the simplest and most commonly used sampling methods. In this method, each and every unit in the population has a equal and independent chance of being included in the sample and each possible sample has an equal chance of being selected. Let us consider an example of simple random sampling. Suppose that a population consists of four units, two males and two females: M1, M2 and F1, F2. If we select samples of size 2, all possible samples are ${}^4C_2 = 6$.

The samples

1. M1M2

2. M2F1

3. F1F2

4. M1F1

5. M2F2

6. M1F2

$$P(M1M2) = \frac{1}{6}$$

$$P(M2F1) = \frac{1}{6}$$

$$P(M1F1) = \frac{1}{6}$$

$$P(M2F2) = \frac{1}{6}$$

$$P(M1F2) = \frac{1}{6}$$

$$P(F1F2) = \frac{1}{6}$$

Thus probability of selecting each sample is equal.

Now we consider the individual probabilities

$$P(M1) = P(M1M2) + P(M1F1) + P(M1F2)$$

$$= \frac{3}{6} = \frac{1}{2}$$

$$P(M2) = P(M1M2) + P(M2F1) + P(M2F2)$$

$$= \frac{3}{6} = \frac{1}{2}$$

$$\text{Similarly, } P(F1) = \frac{3}{6} = \frac{1}{2}$$

$$P(F2) = \frac{3}{6} = \frac{1}{2}$$

Thus, every unit in the population has an equal probability of being selected in the sample.

Using the random number table to select a simple random sample

The most commonly used method of selecting a random sample is to use the table of random numbers. Suppose our population consists of 500 units and we have to select a sample of size 12. In the random number tables, the digits 0 to 9 have equal chance of appearing in a particular position. The steps of selecting the sample are as follows:

1. We choose any three columns anywhere in the random number table.
2. Now we move downwards-selecting 3 digit numbers less than 500.

Consider the following random numbers taken from a random number table.

12135	65186	86886
07369	49031	45451
70387	53186	97116
93451	53493	56442
74077	66687	45394
73627	54287	42596
51353	56404	74106
46426	12855	48497
57126	99010	29015
37997	89034	79788
41283	42498	73173
76374	68251	71593
51668	47244	13732
17698	32685	24490
12448	00902	07263
52515	93269	61210
43501	10248	34219
45279	19382	82151
11437	98102	58168
85183	38161	22848

If we start from the 1st three columns and move downwards the 1st number is 121. We select this number and the 121st unit in the sample is included in the population. The next number is 73 thus this unit is also included in the sample. The next number is 703 which is greater than 500. So we reject this number as well as the next number, which is 934. Proceeding similarly our sample of size 10 consists of the units numbered

121, 73, 464, 379, 412, 176, 124, 435, 452, 114, 128 and 424

It is to be noted that one can start anywhere in the random number table and move vertically or horizontally, but the same pattern needs to be followed in the selection of a single sample.

While selecting the sample there are two possibilities viz.

- (i) Sampling with replacement
- (ii) Sampling without replacement

(i) Sampling with replacement

In this case, each unit selected is replaced in the population before drawing the next unit. Suppose there are 100 balls in a box. In the first case the probability of selecting a unit = $\frac{1}{100}$. After a unit is selected, it is replaced in the box. Thus,

Probability of selecting a unit in the second draw = $\frac{1}{100}$ and so on.

In other words, each unit has equal probability of being selected at each draw.

(ii) Sampling without replacement

In this case, a unit once selected is not replaced in the population. Thus, the probability of selecting a unit varies at each draw. For example if we have to choose 3 units out of 100, the probabilities at the three successive draws are $\frac{1}{100}$, $\frac{1}{99}$ and $\frac{1}{98}$.

6.3.1.2 Stratified Random Sampling

In stratified random sampling, the universe is divided into mutually exclusive sub groups or strata. These strata could be on the basis of geographical area, different age groups, gender etc. A simple random sample is then selected from each strata or sub group. Thus, in stratified random sampling, the entire population of size N is divided into strata of sizes N_1, N_2, \dots, N_k (say k strata) such that the strata are homogenous within themselves and heterogeneous among themselves

$$N = N_1 + N_2 + \dots + N_k.$$

To select a random sample of size n , we choose by simple random sampling n_1 units from the 1st strata, n_2 units from the second strata, and n_k units from the k^{th} strata such that

$$n = n_1 + n_2 + \dots + n_k.$$

Pictorially it may be represented as follows:

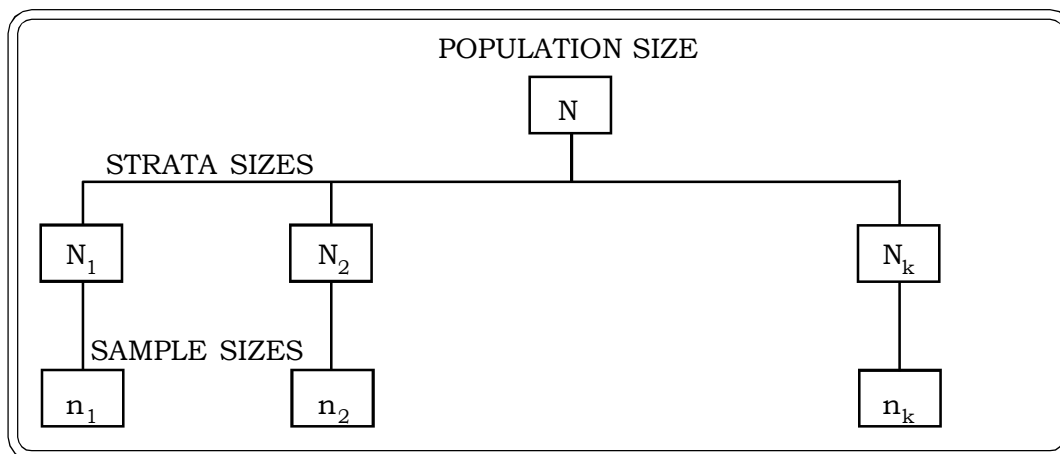


Figure 6.2

Stratified Random Sampling

As an example, suppose a market research organization is doing a survey on a new mobile phone, which has just been launched in the market. They want the sample to represent all age groups. For this, they may stratify the population into different age groups. For example the strata might be 15 years – 25 years, 26 years – 35 years, 36 years – 45 years. And then simple random samples may be drawn from each of these three strata. Each of these samples will be representative of the strata they are drawn from. As another example, consider a company which has 200 part time workers and 1000 full time staff. The company is interested in studying the productivity of both types of workers. Each group needs to be well represented so we can consider one strata as that of the part time workers and the second strata as that of the full time staff. Thus

$$N_1 = 200 \text{ and } N_2 = 1000$$

Next we take samples from these two strata. Suppose we need to choose a sample of size 100. The next step is to decide how many part time and how many full time staff to select so that the sample accurately reflects the proportions of the two groups being studied.

There are two ways in which this can be done viz:

(i) Proportional Stratification

(ii) Disproportional Stratification

(i) In Proportional stratification sample sizes will be chosen by the rule

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \frac{n}{N}$$

In this case

$$\frac{n_1}{N_1} = \frac{n}{N}$$

$$n_1 = \frac{nN_1}{N}$$

$$= \frac{100 \times 200}{1200} = 16.67$$

$$\cong 17$$

i.e. a random sample of size 17 is chosen from the first strata

$$\frac{n_2}{N_2} = \frac{n}{N}$$

$$n_2 = \frac{nN_2}{N}$$

$$= \frac{100 \times 1000}{1200}$$

$$\cong 83$$

A random sample of size 83 is selected from the second strata.

The general rule for proportional stratification for k strata's:

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \dots = \frac{n_k}{N_k} = \frac{n}{N}$$

(ii) In **disproportional stratification** we may select samples from various strata by considering different weightage to different strata. In this example if we feel that the full time staff needs more representation then we may select a higher percentage of the sample from this strata.

6.3.1.3 Systematic Sampling

In systematic sampling, the units in the population are selected at regular intervals for example once in a day, a unit after every 2 lots of production, etc. In this method, only the 1st unit is selected at random. The rest of the units are selected according to a pattern depending on a factor $k = \frac{n}{N}$ which is also called the sampling ratio. Suppose our population consists of 100 units and we have to draw a sample of size 10. Thus $k = \frac{1}{10}$. The units are arranged as follows in groups of 10.

Table 6.1
Systematic Sampling

I	II	III	IV	V	VI	VII	VIII	IX	X
1	11	21	31	41	51	61	71	81	91
2	12	22	32	42	52	62	72	82	92
3	13	23	33	43	53	63	73	83	93
4	14	24	34	44	54	64	74	84	94
5	15	25	35	45	55	65	75	85	95
6	16	26	36	46	56	66	76	86	96
7	17	27	37	47	57	67	77	87	97
8	18	28	38	48	58	68	78	88	98
9	19	29	39	49	59	69	79	89	99
10	20	30	40	50	60	70	80	90	100

A number from 1 and 10 is now selected at random. Suppose it is 3. We select the 3rd unit and every 10th unit thereafter. Thus the sample will consist of the following units

3, 13, 23, 33, 43, 53, 63, 73, 83, 93.

In practice, suppose we wish to check if the metro rail is running on time or not. From a total of 80 trains suppose that we want to sample 10 trains. The sampling fraction is $\frac{80}{10} = 8$ so every

8th train passing through a station is sampled after a random starting point between 1 and 8 is chosen. If the number is 2, the trains selected are 2, 10, 18, 26, 34, 42, 50, 58, 66 and 74th trains.

6.3.1.4 Cluster sampling

In cluster sampling, the population is sub – divided into certain groups or clusters. We first select a random sample of clusters and from these selected clusters, random units are then selected for study. For example, to study average household income in say Maharashtra, clusters. could be the various districts. A sample of districts is first selected and then households are again randomly chosen from the selected districts. Cluster sampling is generally used when the population has natural groupings, usually in terms of geographical areas.

6.3.1.5 Multistage Sampling

This sampling procedure consists of selecting a sample in a series of steps by dividing the population into several successive stages. For example, if we wish to survey the reading habits of school children, the first stage units would be a sample of cities, the second stage units would be schools in each city and the third stage unit could be classes in each school and the fourth stage units could be children in each class.

6.3.2 Non-Probability Sampling Methods

These methods are such that the sample units are not selected on the basis of any probability. These methods are non random in nature.

A few non-probability sampling methods are

- (i) Judgement Sampling
- (ii) Convenience Sampling
- (iii) Quota Sampling

6.3.2.1 Judgement Sampling

In this type of sampling, the investigator decides which units to include or exclude in the sample. For example, suppose in a tyre-manufacturing factory it is required to select 10 very good pieces for some studies. The production manager will set certain parameters and choose the 10 units at his own discretion. A probability sampling method would not solve the purpose here. Judgement sampling though very simple and convenient has one major drawback. It may be influenced by personal bias of the investigator. Any preconceptions of the researcher may get reflected in the sample. However, this method is quite suitable in exploratory surveys like pre-testing of questionnaire.

6.3.2.2 Convenience Sampling

This sampling is also by taking into consideration convenience of the investigator. For example to conduct a survey of customer satisfaction in a retail outlet, the investigator might choose to stand a particular point according to his/her own convenience and question people or for a survey related to customers. In a bank, the investigator may decide to conduct the survey with the first 50 customers (say). This method is not very efficient. However, sometimes it may be the only recourse. It is also used quite a lot in pilot surveys before say, launching of a product in the market.

6.3.2.3 Quota Sampling

In this method, certain parameters are pre- decided and the investigator is given the authority to apply his discretion in certain other cases. For example, to study eating habits of school children and college students it may have been pre – decided that 100 school & 150 college students would be in the sample. The final choice of the 100 school students & 150 college students is left to the discretion of the investigator. As another example, consider a quota sample of size 20 based on age groups.

Table 6.2
Quota Sampling

Age	Sample Size
20 - 30	4
30 - 40	6
40 - 50	7
50 and above	3

Thus, sample sizes in each group are fixed. However, investigator has the freedom to choose the units in each age group according to his/her convenience.

6.3.3 Sampling and Non – Sampling Errors

Sampling errors arise from the fact that a sample has been used to study the population. These errors are generally not present in a complete census as they are associated with the process of selecting a sample. A sampling error is the difference between the estimate obtained from sampling and the true value for the entire population. Sampling errors may be due to faulty selection of the sample, improper data collection method, incorrect methodology of analyzing the data and so on. Sampling errors tend to decrease as the sample size increases as shown in the following figure 6.3

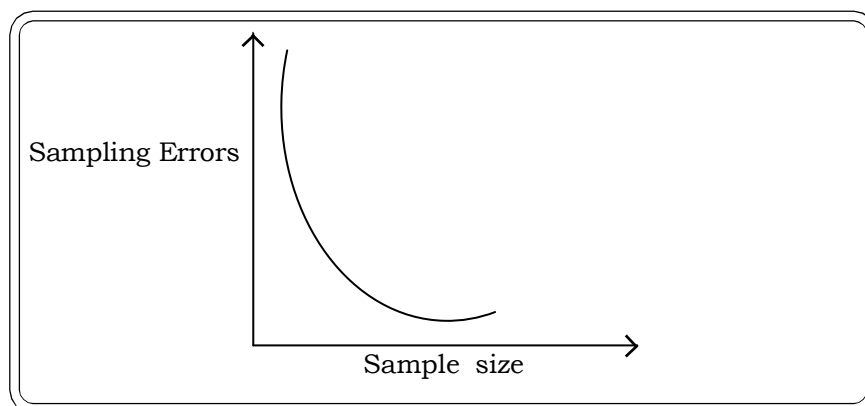


Figure 6.3

Relation between Sampling Error & Sampling Size

Non – sampling errors are errors arising during the course of all survey activities other than sampling. They exist both in sample surveys as well as censuses. Errors due to non-response, lack of trained investigators, errors in estimation and analysis in data processing, etc. are examples of non – sampling errors. Unlike sampling errors, non – sampling errors tend to increase with the increase in sample size as shown in the graph in figure 6.4.

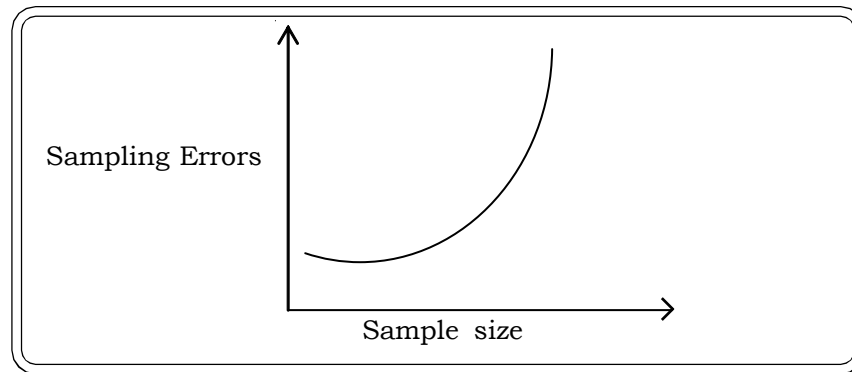


Figure 6.4

Relation between Non- Sampling Errors & Sampling Size

6.4 SAMPLING DISTRIBUTIONS

In section 6.2, we discussed the concepts of a parameter and a statistic. The discussion was concluded with the statement that the statistic, which is a function of sample values, is a random variable. Now, the next question arises that if it is a random variable it will take different values and hence what are the probabilities that it assumes certain values? This theoretical probability distribution of the statistic is known as the sampling distribution. In particular we discuss example of two statistics viz. arithmetic mean and proportion. The probability distributions, which these two statistics follow are referred to as their sampling distribution.

Let us consider an example. Suppose an advertising company wants to find out the proportion of teenagers who use a particular brand of shampoo. Since all teenagers cannot be interviewed, the company decides to question 100 teenagers and we use this data to make a final conclusion. Suppose the sample data reveals that the proportion in the sample is 60%. Now, it is quite possible that another sample of 100 may give the proportion as 70% and yet another may give a different proportion. All these values taken together will describe the distribution of sample proportions.

In general, one may conceive of all possible samples of a given size that can be drawn from the population being studied. For each of these samples we compute the statistic of interest say for example the mean or the proportion. The sampling distribution of this statistic is the probability distribution of all these values.

We now illustrate a sampling distribution with the help of the following example.

Consider a firm with 4 branch offices each employing 4, 8, 6 and 10 people. This data can be arranged as follows:

Table 6.3
Employees in a firm

Firm Number	No. of Employees (x)
1	4
2	8
3	6
4	10

The four branches constitute the population. We next calculate the following quantities from this population.

Population mean

$$\mu = \frac{\Sigma X}{n} = \frac{4+8+6+10}{4} = \frac{28}{4} = 7$$

Population variance

$$\begin{aligned}\sigma^2 &= \frac{1}{n} \Sigma (X - \mu)^2 = \frac{(4-7)^2}{4} + \frac{(8-7)^2}{4} + \frac{(6-7)^2}{4} + \frac{(10-7)^2}{4} \\ &= \frac{20}{4} = 5\end{aligned}$$

Population standard deviation

$$\sigma = 2.24$$

Now, consider all possible samples of size 2 without replacement from this population i.e. ${}^4C_2 = 6$ possible samples. The table given below lists all the 6 possible samples along with the mean number of employees of each sample.

Table 6.4

Mean Employees of all possible samples of size 2 from a population

Sample Number	Firm Number	Number of Employees	Mean Employees
1	(1, 2)	(4, 8)	$\bar{x}_1 = 6$
2	(1, 3)	(4, 6)	$\bar{x}_2 = 5$
3	(1, 4)	(4, 10)	$\bar{x}_3 = 7$
4	(2, 3)	(8, 6)	$\bar{x}_4 = 7$
5	(2, 4)	(8, 10)	$\bar{x}_5 = 9$
6	(3, 4)	(6, 10)	$\bar{x}_6 = 8$

An examination of the mean employees shows considerable difference among the six samples and also as compared to the population mean, most of the means differ.

However, on calculating the grand mean or the mean of the mean employees, we get

$$\text{Grand Mean} = \frac{6+5+7+7+9+8}{6} = 7$$

which is same as the population mean. Thus, mean of the sample means coincides with the population mean.

Now, considering the values of the mean employees of the samples as a random variable, we can define its probability distribution as follows:

Table 6.5
Probability Distribution of Mean Employees

Mean Employees	Frequency	Probability
5	1	$\frac{1}{6} = 0.17$
6	1	$\frac{1}{6} = 0.17$
7	2	$\frac{2}{6} = 0.32$
8	1	$\frac{1}{6} = 0.17$
9	1	$\frac{1}{6} = 0.17$
		1

Thus, the probability distribution of the sample means as shown above is known as the sampling distribution of the statistic - sample mean. Similarly, it is possible to define sampling distribution for other sample statistics.

6.4.1 The Central Limit Theorem (CLT)

The central limit theorem is one of the most significant results in the field of probability and has wide ranging implication and applications. It is the foundation for many statistical procedures especially inferential statistics and also quality control methods. Stated simply, it says that the distribution of an average tends to be normal, regardless of the distribution from which the average is computed, provided the sample size is large.

Also, this normal distribution will have the same mean as the parent distribution and variance equal to the variance of the parent distribution divided by the sample size.

Statement of the Central Limit Theorem

Let x_1, x_2, \dots, x_n be a random sample of observations drawn from any population with mean μ and variance σ^2 . Define

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Then, the distribution of \bar{x} can be approximated by a normal distribution with mean and variance $\frac{\sigma^2}{n}$, provided n -the sample size is large.

Symbolically,

$$\bar{x} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right), \text{ provided } n \text{ is large}$$

where $\mu \rightarrow$ mean of the parent distribution.

$\sigma \rightarrow$ Variance of the parent distribution

$n \rightarrow$ Size of the sample

Consider the following example:

The following distribution is a uniform distribution in the range [0,1]

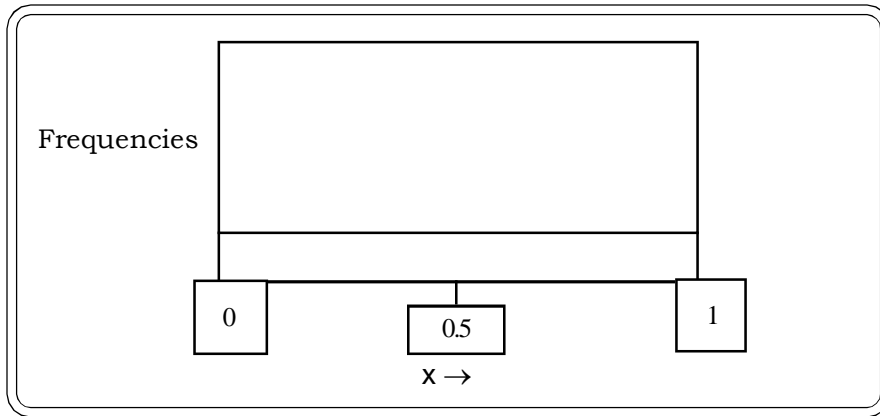


Figure 6.5

Uniform Distribution

Now, two samples are drawn from this distribution at random and their average is computed. Another sample of size two is selected & the average computed. This process is repeated a number of times. The distribution of the average for samples of size 2 is shown in the figure below.

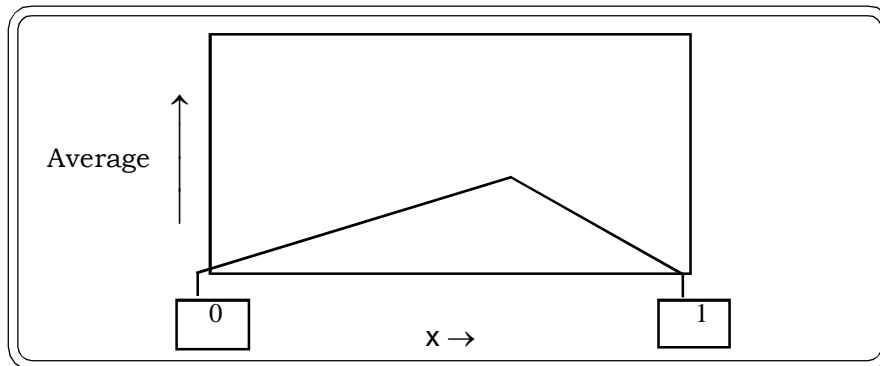


Figure 6.6

Distribution of the Average for two Samples

Proceeding similarly, the distribution of the average by repeatedly taking about 30 samples of size 2 from the parent population, which is uniform, will produce the following graph (figure 6.7) which is the very familiar bell shaped normal curve. Thus, as the sample size increases, we could see the transition of a uniform distribution to a normal distribution.

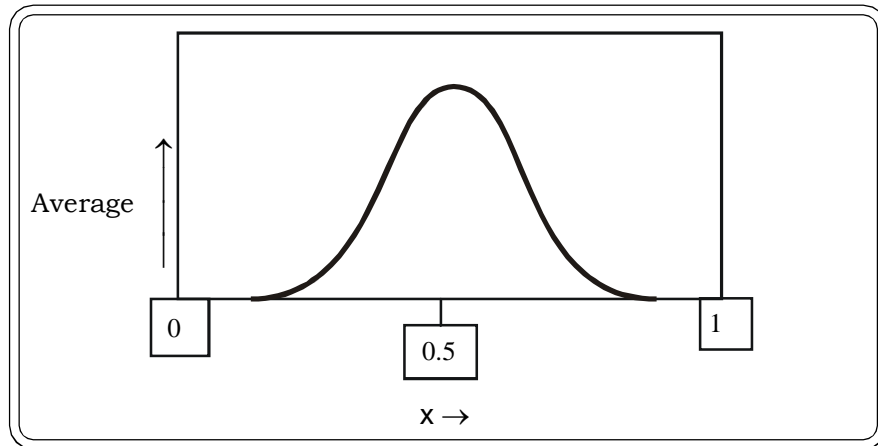


Figure 6.7

Uniform Distribution Approaching Normal Distribution as $n \rightarrow \infty$

This in essence is the central limit theorem. For non-normal parent distributions also, the average can be approximated by a normal distribution, as the sample size increases.

Alternatively, by converting to the Z - scale, the central limit theorem can be stated as

The distribution of

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

is approximately standard normal, when the sample size is large.

Thus, $Z \sim N(0, 1)$

Remarks:

(i) We have been talking about the sample size being large throughout our discussion on CLT. The question arises how large a sample would be considered appropriate.

If the parent population is normally distributed, the distribution of the average will be normal for any sample size.

If the parent distribution is symmetric, a reasonably good approximation can be obtained for $n = 10$

If the parent population is not symmetric, then the convergence to normality will be much more slower and samples of size 30 or more will be necessary.

(ii) If the population is finite and samples of size n are drawn without replacement, then the

standard error of the mean is given by $\sigma_1 = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

The term $\sqrt{\frac{N-n}{N-1}}$ is called the finite correction factor.

6.4.2 Sampling Distribution of the Mean

This is essentially the Central Limit Theorem. The sample mean \bar{x} of a set of observations x_1, x_2, \dots, x_n follows a normal distribution with

$$E(\bar{x}) = \mu$$

and $V(\bar{x}) = \frac{\sigma^2}{n}$, provided n is large

where μ = mean of the parent distribution

σ = Variance of the parent distribution

n = Sample size

We now look at examples of application of the CLT.

Example 6.1: A sample consists of 100 units. Mean of the population is 50 and standard deviation 5. A sample of size 10 is drawn from this population. Find the mean and standard deviation of the distribution of the sample mean.

Solution:

Given $\mu = 50$

$$\sigma = 10$$

$$n = 100$$

$$E(\bar{x}) = 50$$

$$sd(\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{10}} = 4.46$$

Thus, mean of sample mean = 50

And standard deviation of sample mean = 4.46

Example 6.2: Suppose that a population has the following distribution:

x	-2	0	4	8
p(x)	0.3	0.2	0.3	0.2

- (i) Find the mean and variance of this population.
- (ii) Find the distribution of the sample mean if a sample of 100 observations is drawn from the population.
- (iii) Calculate the following probabilities:
 - $P(1.36 < \bar{x} < 2.8)$
 - $P(\bar{x} > 2.3)$
 - $P(\bar{x} < 1.4)$

Solution:

$$\begin{aligned}
 \text{(i) } \mu &= \sum xp(x) \\
 &= (-2)(0.3) + 0(0.2) + 4(0.3) + 8(0.2) \\
 &= -0.6 + 0 + 1.2 + 1.6 \\
 &= 2.2
 \end{aligned}$$

$$\begin{aligned}
 \sigma^2 &= \sum x^2 p(x) - \mu^2 \\
 &= [4(0.3) + 0(0.2) + 16(0.3) + 64(0.2)] - (2.2)^2 \\
 &= [1.2 + 4.8 + 12.8] - 4.84 \\
 &= 18.8 - 4.84 \\
 &= 13.96
 \end{aligned}$$

Thus, population mean = 2.2

Population variance = 13.96

Population standard deviation (σ) = 3.73

(ii) According to the central limit theorem, the sample mean \bar{x} will follow a normal distribution

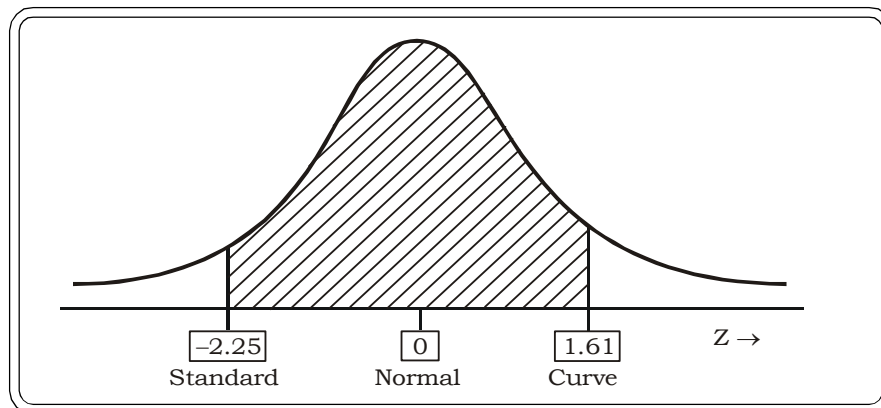
with mean μ i.e. 2.2 and standard deviation $\left(\frac{\sigma}{\sqrt{n}}\right)$

$$\text{Thus } \frac{\sigma}{\sqrt{n}} = \frac{3.73}{10} = 0.373$$

Thus $\bar{x} \rightarrow N(2.2, 0.373)$

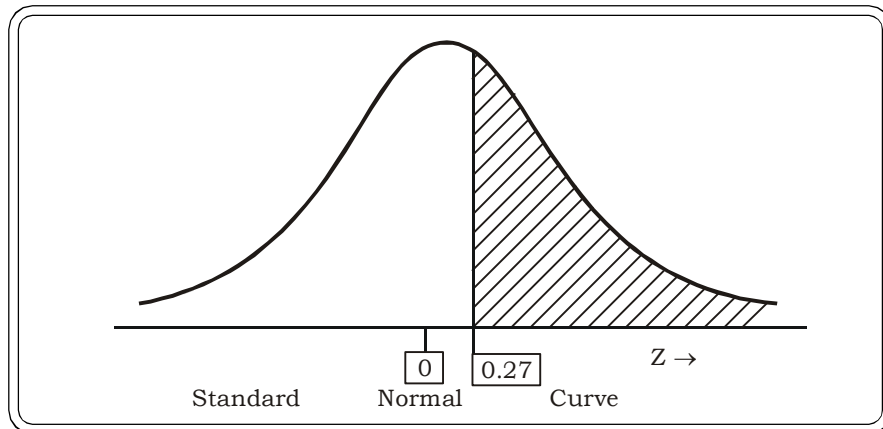
(iii) Standardizing by using the distribution of \bar{x}

$$\begin{aligned}
 &P\left(\frac{1.36 - 2.2}{0.373} < Z < \frac{2.8 - 2.2}{0.373}\right) \\
 &= P(-2.25 < Z < 1.61) \\
 &= P(0 < Z < 1.61) + P(0 < z < 2.25) \\
 &= 0.4463 + 0.4878 \\
 &= 0.9341
 \end{aligned}$$



Thus $P(1.36 < \bar{x} < 2.8) = 0.9341$

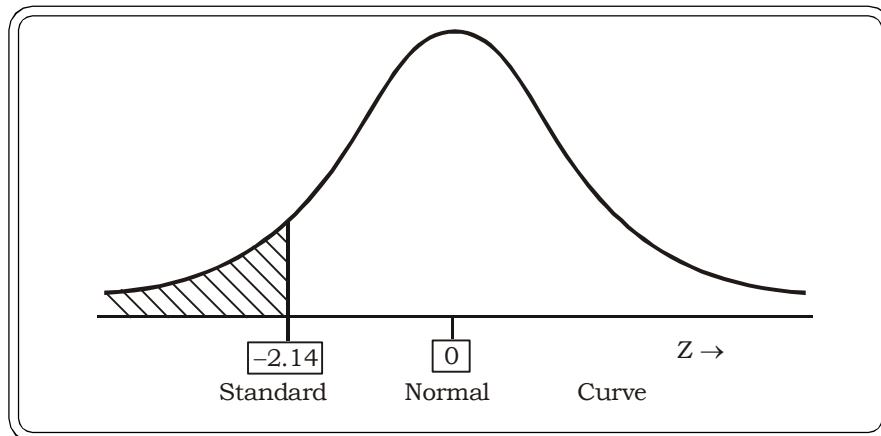
$$\begin{aligned}
 P(\bar{x} > 2.3) &= P\left(Z > \frac{2.3 - 2.2}{0.373}\right) \\
 &= P(Z > 0.27)
 \end{aligned}$$



$$\begin{aligned}
 &= 0.5 - P(0 < Z < .27) \\
 &= 0.5 - 0.1064 \\
 &= 0.3936
 \end{aligned}$$

Thus, $P(\bar{x} > 2.3) = 0.3936$

$$P(\bar{x} < 1.4) = P\left(Z < \frac{1.4 - 2.2}{0.373}\right) = P(Z < -2.14)$$



$$\begin{aligned}
 &= 0.5 - P(0 < Z < 2.14) \\
 &= 0.5 - 0.4838 \\
 &= 0.0162
 \end{aligned}$$

Example 6.3: A bank has 300 employees on its payroll. The average annual salary of the 300 employees is estimated to be Rs.4, 00,000 with a standard deviation of Rs.1, 00, 000. In a sample of 100 employees find the probability that the average salary will be less than Rs.3, 75,000?

Solution:

Let $X \rightarrow$ denote the salary of bank employees

Let μ = average annual salary of the 300 bank employees

σ = standard deviation of the same

Then μ = Rs. 4, 00, 000

σ = Rs. 1, 00,000

Sample Size $n = 100$

We have to find

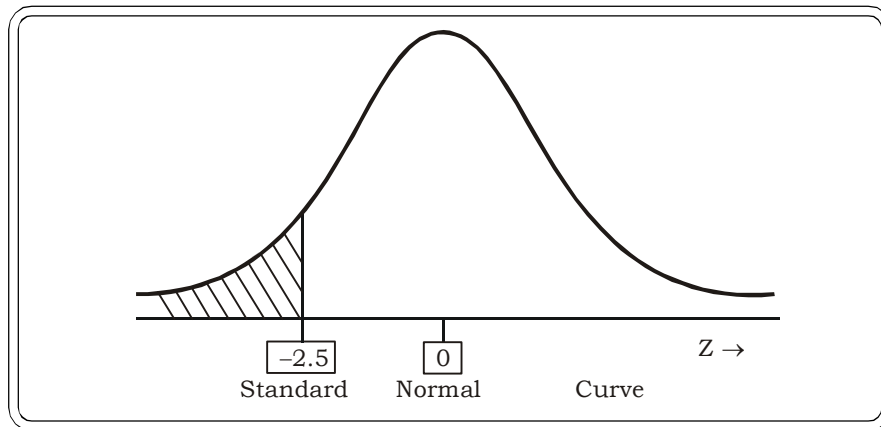
P (average salary of 50 employees in the sample < 3,75,000)

$$= P(\bar{x} < 3,75,000)$$

$$= P\left(Z < \frac{3,75,000 - 4,00,000}{\frac{1,00,000}{\sqrt{100}}}\right)$$

$$= P\left(Z < \frac{25,000}{10,000}\right)$$

$$= P(Z < -2.5)$$



$$= 0.5 - P(0 < Z < 2.5)$$

$$= 0.5 - 0.4938$$

$$= 0.0062$$

Thus the probability that the average salary is less than Rs.3, 75,000 is 0.062

Example 6.4: A firm produces light bulbs that are known to be normally distributed with a mean lifetime of 1200 hours and a standard deviation of 210 hours. What is the probability that a simple random sample of 100 bulbs will yield a mean that falls between 1,140 and 1,260 hours?

Solution:

Let $x \rightarrow$ the lifetimes of the bulbs

Mean lifetime of the bulbs (μ) = 1200 hours

Standard deviation of lifetime of bulbs (σ) = 210 hours

Random sample size (n) = 100

By the CLT

The sample mean lifetime (\bar{x}) $\rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

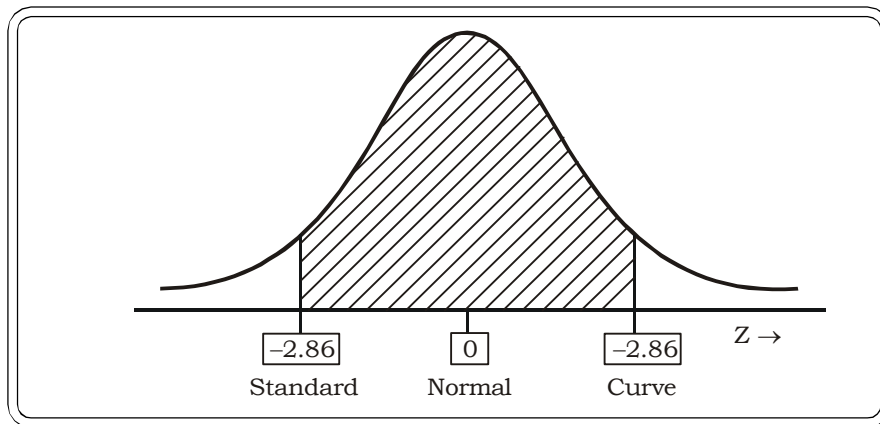
$$\frac{\sigma}{\sqrt{n}} = \frac{210}{\sqrt{100}} = \frac{210}{10} = 21$$

Thus $\bar{x} \rightarrow N(1200, 21)$

We need to find $P(1,140 < \bar{x} < 1260)$

$$P(1,140 < \bar{x} < 1260) = P\left(\frac{1140-1200}{21} < Z < \frac{1260-1200}{21}\right)$$

$$P = (-2.86 < Z < 2.86)$$



$$= 2 P(0 < Z < 2.86)$$

$$= 2(0.4979)$$

$$= 0.9958$$

Thus 99.58% of the bulbs will yield a mean that is between 1,140 and 1,260 hours.

Example 6.5: A manufacturer of knitting yarn has established from past experience that the breaking strength of this yarn is normally distributed with a mean of 12 pounds and standard deviation of 1.8 pounds. What is the probability that a sample size of 49 yield a mean of 12.5 pounds or more?

Solution:

Let x -breaking strength of knitting yarn.

Then $x \rightarrow N(12, 1.8)$

i.e. $E(x) = 12$ pounds

and $V(x) = (1.8)^2$

The sample size (n) = 49

By CLT, the distribution of mean (\bar{x}) will be normal with

$E(\bar{x}) = 12$

$$\text{s.d.}(\bar{x}) = \frac{1.8}{\sqrt{49}} = \frac{1.8}{7} = 0.26$$

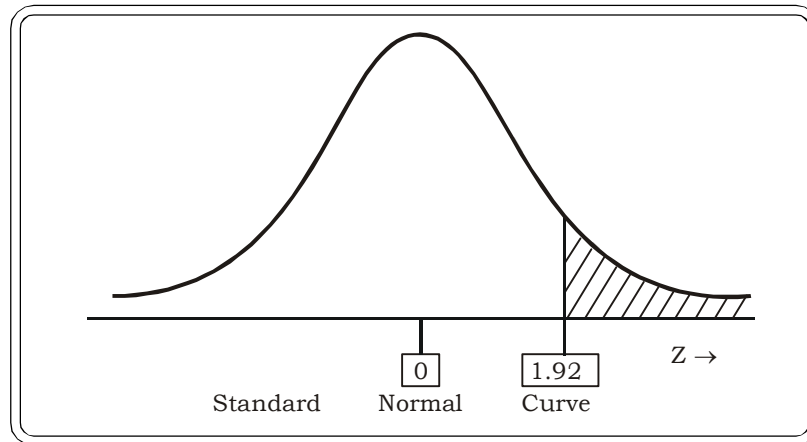
We have to determine

$$P(\bar{x} \geq 12.5)$$

Transforming to the standard scale

$$P\left(\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \geq \frac{12.5 - 12}{0.26}\right)$$

$$= P(Z \geq 1.92)$$



$$= 0.5 - 0.4726$$

$$= 0.0274$$

Thus, the probability that a sample size of 49 yields a mean of 12.5 pounds or more is 0.0274

Example 6.6: It is known that the annual income of a family has a distribution with mean \$12,200 and a standard deviation of \$3,500. If a sample of 64 families is picked at random, find the standard error of the mean family income of the population.

Solution:

Let x – the annual income of a family

\bar{x} - mean family income

Sample size = 64

The standard error of the mean family income

$$= \frac{\sigma}{\sqrt{n}} = \frac{3500}{8} = 437.5$$

Note: Standard Error – The standard deviation of statistic is referred to as the standard error.

Example 6.7: A cigarette – manufacturing company claims that the mean nicotine content in their king size cigarettes is 2 mg and the standard deviation of the nicotine content is equal to 0.3mg. If this claim is valid, what is the approximate probability that a sample of 900 cigarettes will yield a mean nicotine content exceeding 2.02 mg?

Solution:

Let $X \rightarrow$ Nicotine content in the cigarettes

Mean nicotine content of their king size cigarette (μ) = 2mg

And standard deviation (μ) = 0.3mg

Sample Size (n) = 900 cigarettes

We have to find the probability that the mean nicotine content exceeds 2.02 mg i.e.

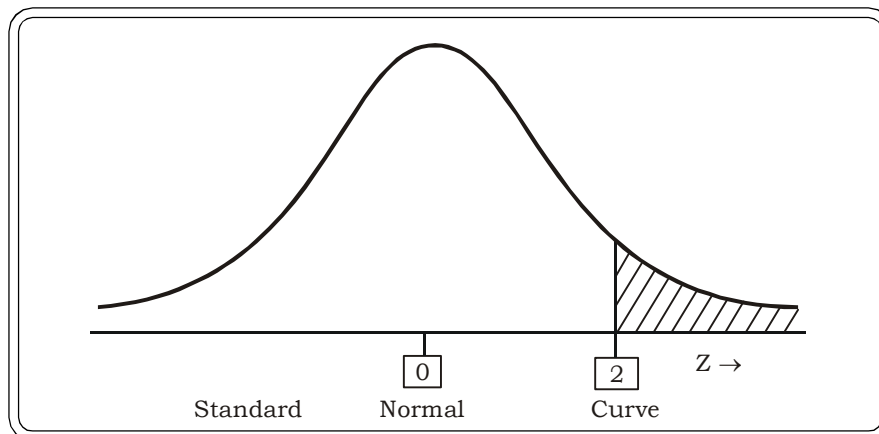
$$P(\bar{x} > 2.02)$$

$$\text{Standard error of the mean} = \frac{0.3}{\sqrt{900}} = 0.01$$

i.e. the mean follows a normal distribution with mean nicotine content 2 mg & standard deviation 0.03 mg

Converting to the standard scale,

$$P(\bar{x} > 2.02) = P\left(\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} > \frac{2.02 - 2}{0.01}\right) = P(Z > 2)$$



$$= 0.5 - 0.4772$$

$$= 0.0228$$

Example 6.8: The weight of chocolate bars packed in a box is a random variable with mean weight 16 gms and standard deviation 0.6 gms. If the boxes contain 36 chocolates bars, what is the probability that the mean weight of a randomly picked box will be over 585 gms.

Solution:

Let $X \rightarrow$ The weight of chocolate bars packed in a box.

$$\mu = 16 \text{ gms.}$$

$$\sigma = 0.6 \text{ gms}$$

$$n = 36$$

\bar{x} - mean weight of the chocolate bar

We have to find

$$P(\text{mean weight of a chocobar in the box} > \frac{585}{36})$$

$$= P(\bar{x} > 16.25 \text{ gms})$$

The distribution of \bar{x} is normal with mean = 16 gms.

Standard Deviation = 0.6 gms

Thus

$$\begin{aligned} P(\bar{x} > 16.25 \text{ gms}) &= P\left(Z > \frac{16.25 - 16}{0.6/6}\right) \\ &= P(Z > 2.5) \\ &= 0.0062 \end{aligned}$$

6.4.3 Sampling Distribution of the Proportion

Consider the following situation.

Suppose 5% of the tyres produced by a machine are defective. A sample of 100 tyres is chosen and we want to determine the probability that the proportion of defectives will exceed say 1%.

Thus, to analyze qualitative data we need to relate the sample proportion to the population proportion. The sampling distribution of the proportion is the distribution of proportions of all possible random samples of size n .

Let X – number of units in population possessing a certain attribute.

N – size of the population.

Then $P = \frac{X}{N}$ is the proportion of units in the population possessing a certain attribute

Let n – sample size

x – number of units in the sample possessing the attribute.

Then $p = \frac{x}{n}$ is sample proportion of units possessing the attribute.

Using the central limit theorem, if the sample size is large, the distribution of the sample proportion $p = \frac{x}{n}$ is a normal distribution (provided p is not very close to 0 or 1) with mean P and

$$\text{variance } \frac{P(1-P)}{n} = \frac{\frac{X}{N}\left(1 - \frac{X}{N}\right)}{n}$$

Converting to the z scale:

$$Z = \frac{\frac{x}{n} - P}{\sqrt{\frac{P(1-P)}{n}}} \text{ follows a standard normal distribution, provided } n \text{ is large.}$$

Example 6.9: It is believed that 40% of the people favor capital punishment. If 400 persons are interviewed at random, what is the probability that the proportion of individuals in the sample who favor capital punishment will exceed 0.43?

Solution:

The population proportion of people who favor capital punishment = 0.40

Random sample size = 400

The sample proportion will follow a normal distribution with mean 0.40 and standard deviation

$$\sqrt{\frac{0.40(1-0.40)}{400}} = \sqrt{.0006} = 0.024$$

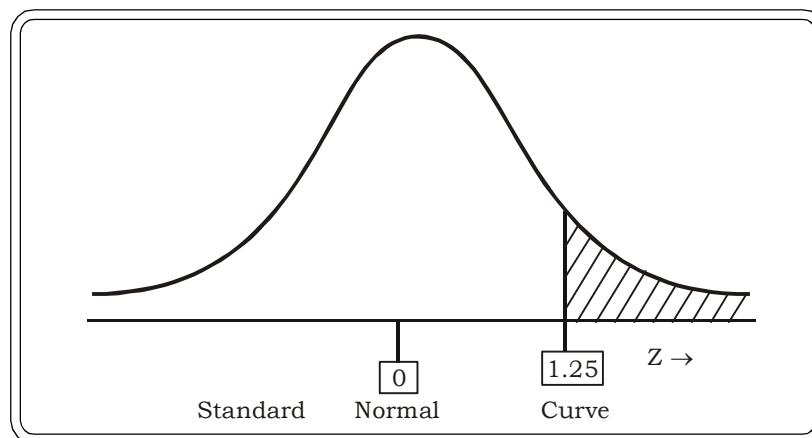
we have to find

$P(p > 0.43)$ i.e. Probability that the sample proportion exceeds 0.43

Converting to the Z scale

$$P\left(\frac{p - P}{\sqrt{\frac{P(1-P)}{n}}} > \frac{0.43 - 0.40}{0.024}\right)$$

$$= P(Z > 1.25)$$



$$= 0.5 - P(0 < Z < 1.25)$$

$$= 0.5 - 0.3944$$

$$= 0.1056$$

Thus, probability that the proportion of individuals in the sample who favor capital punishment will exceed 0.43 is 0.1056

Example 6.10: It is found that 70% of estimated students of a certain University feel that the American military forces should be withdrawn from Iraq. What is the probability that in a sample of 64 students interviewed, less than 60% will favor withdrawal of American military forces from Iraq?

Solution:

P = The population proportion of students who feel that the American military forces should be withdrawn. = 0.70

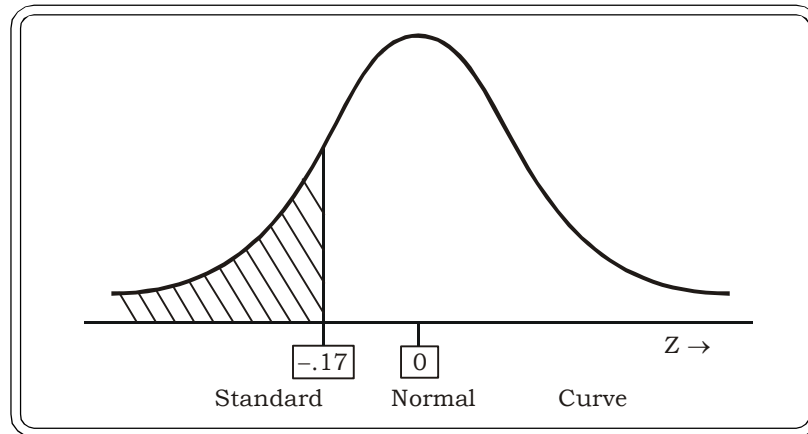
n = The sample size = 64

$$\text{Standard error of the sample proportion} = \sqrt{\frac{0.70(1-0.70)}{64}} = \sqrt{.0033} = 0.057$$

Thus $P(p < 0.60)$

$$= P\left(Z < \frac{0.60 - 0.70}{0.057}\right)$$

$$= P(Z < -0.175)$$



$$= 0.5 - 0.0714 = 0.4286$$

Example 6.11: A car company has estimated that 5% of all new cars produced by them are recalled due to defects. A random sample of 100 cars is selected for inspection. What is the probability that between 4 to 6 cars will be defective in this sample?

Solution:

Population proportion of defective cars = 0.05

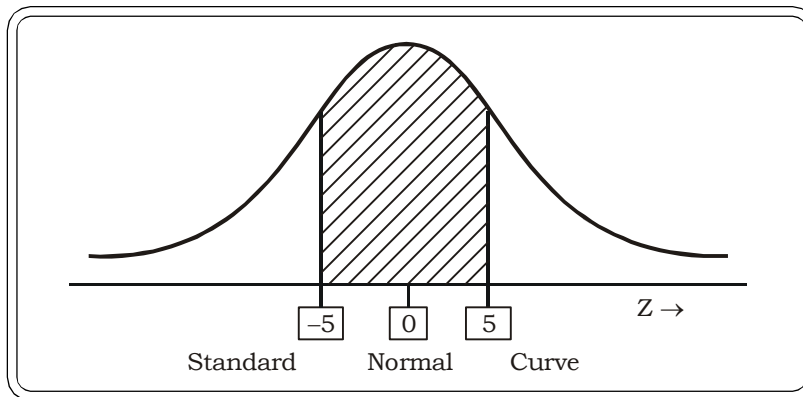
The sample size = 100

$$\begin{aligned} \text{Standard error of sample proportion} &= \sqrt{\frac{0.05(1 - 0.05)}{100}} = \sqrt{0.000475} \\ &= 0.02 \end{aligned}$$

Thus, the mean of the sample proportion = 0.05 and the standard deviation of the sample proportion = 0.02

By CLT, the sampling distribution of proportion of defective cars is normal

$$\begin{aligned} \text{To find } P\left(\frac{4}{100} < p < \frac{6}{100}\right) \\ &= P(0.04 < p < 0.06) \\ &= P\left(\frac{.04 - .05}{0.02} < Z < \frac{.06 - .05}{0.02}\right) \\ &= P(-0.5 < Z < 0.5) \end{aligned}$$



$$\begin{aligned}
 &= 2 P (0 < Z < 0.5) \\
 &= 2 \times (0.1915) \\
 &= 0.383
 \end{aligned}$$

Thus the probability that between 4 to 6 cars will be defective in this sample is 0.383.

Example 6.12: A bank has determined that 60% of their customers respond to initial requests for confirmation of their account balances. If a simple random sample of 64 customers is sent requests for confirmation, what is the probability that 50% or more will respond?

Solution:

P = Population proportion of customers who respond to initial requests for confirmation of their account balances = 0.60

Random sample size (n) = 64

To find $P (p > 0.50)$

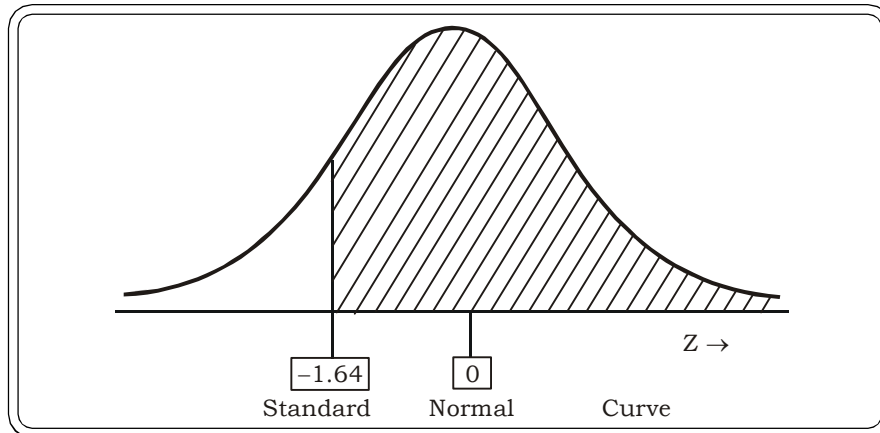
The standard error of the sample proportion : $\sqrt{\frac{P(1-P)}{n}}$

$$= \sqrt{\frac{0.6(1-0.6)}{64}} = \sqrt{.00375} = 0.061$$

Thus, sample proportion (p) of customers who respond to initial request for confirmation follows a normal distribution with mean = 0.60

and standard deviation = 0.061

$$\text{Thus } P (p > 0.50) = P\left(Z > \frac{0.50 - 0.60}{0.061}\right) = P(Z > -1.64)$$



$$\begin{aligned}
 &= 0.5 + P(0 < Z < 1.64) \\
 &= 0.5 + 0.4495 \\
 &= 0.9495
 \end{aligned}$$

The probability that 50% or more will respond is 0.9495.

Example 6.13: A manufacturer of printed circuit boards (PCB's) has determined that 3% of the PCB's he produces are defective. In a random sample of 500 PCB's, what is the probability that the proportion of defective PCB's is between 0.025 and 0.045?

Solution:

$$P = 0.03 \text{ (Proportion of defective PCB's)}$$

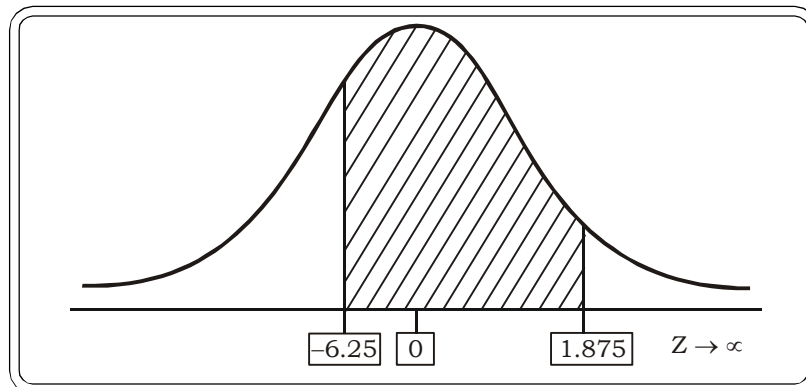
$$n = 500 \text{ PCB's}$$

To find $P(0.025 < p < 0.045)$

$$S.E.(p) = \sqrt{\frac{(.03)(1-.03)}{500}} = \sqrt{.0000582} = .008$$

Thus $p \rightarrow N(.03, (.008)^2)$, by the Central Limit Theorem

$$P(0.025 < p < 0.045) = P\left(\frac{0.025 - 0.03}{0.008} < Z < \frac{0.045 - 0.03}{0.008}\right) = P(-0.625 < Z < 1.875)$$



$$\begin{aligned}
 &= P(0 < Z < 0.625) + P(0 < Z < 1.875) \\
 &= 0.2324 + 0.4692 \\
 &= 0.7016
 \end{aligned}$$

The probability that the proportion of defective PCB's is between 0.025 and 0.045 is 0.7016.

Example 6.14: A local bank has 2000 depositors with 40% of these depositors having current as well as savings account. The rest have only current accounts. A random sample of 400 such accounts has been selected. What is the probability that the sample proportion of depositors with both accounts will be between 0.40 and 0.43?

Solution:

Population size (N) = 2000

Population proportion of depositors having current as well as savings accounts (P) = 0.40

Random Sample size (n) = 400

Let p = sample proportion of depositors with both accounts

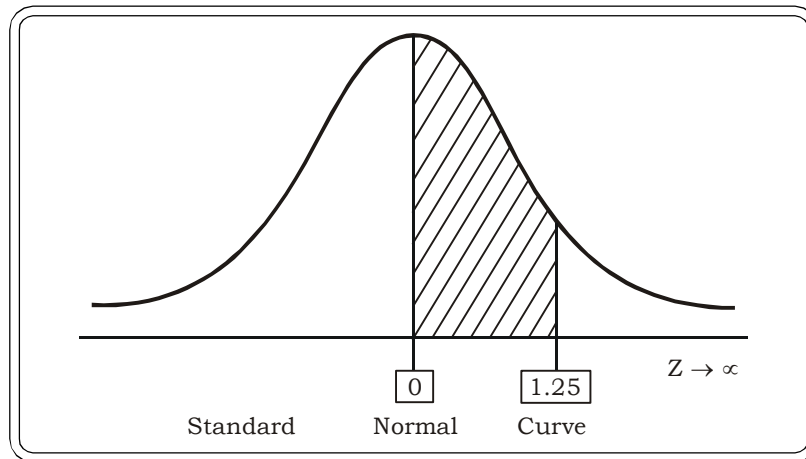
$$\text{Then } p \rightarrow N\left(0.40, \frac{0.40(1-0.40)}{400}\right)$$

$$p \rightarrow N(0.40, 0.024)$$

We have to find $P(0.40 < p < 0.43)$

Converting to the Z scale, (since the distribution of p is normal)

$$\begin{aligned} P(0.40 < p < 0.43) &= P\left(\frac{0.40 - 0.40}{0.024} < p < \frac{0.43 - 0.40}{0.024}\right) \\ &= P(0 < Z < 1.25) \end{aligned}$$



= 0.3944 (from standard normal tables)

Thus, the probability that the sample proportion of depositors with both accounts will be between 0.40 and 0.43 is 0.3944

Example 6.15: A courier company claims that 95 percent of all mail are delivered the very next day if within the country. To test the claim 400 deliveries were selected.

- (i) Find the probability that between 94% to 96% of deliveries will be completed in one day.
- (ii) Also find the probability that more than 98% of the couriers will be delivered by the next day.

Solution:

Given population proportion:

$$P = 0.95$$

$$n = 400$$

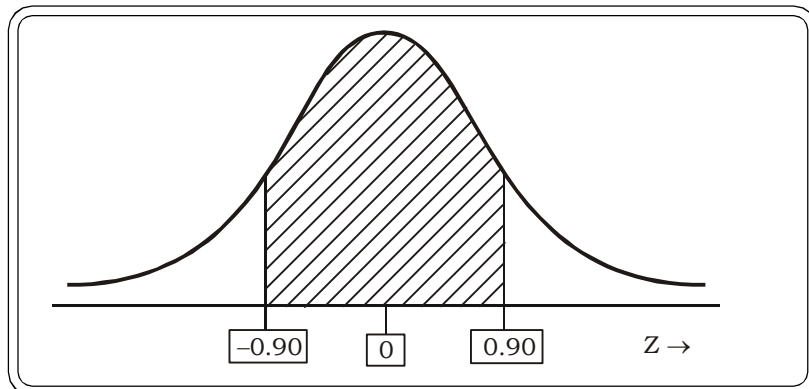
$$\text{Sample standard error} = \sqrt{\frac{0.95(1-0.95)}{400}} = \sqrt{.00012} = 0.011$$

The distribution of the sample proportion is normal with mean = 0.95 and standard deviation = 0.011

Now, we have to find

- (i) $P(0.94 < p < 0.96)$ i.e. probability that 94% to 96% of deliveries would be completed on one day.

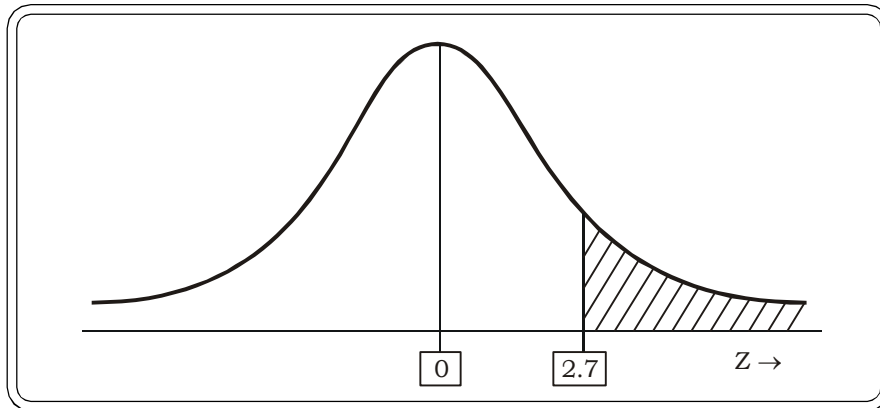
$$\begin{aligned} &= P\left(\frac{0.94-0.95}{0.011} < Z < \frac{0.96-0.95}{0.011}\right) \\ &= P(-0.90 < Z < 0.90) \end{aligned}$$



$$\begin{aligned} &= 2 P(0 < Z < 0.90) \\ &= 2 (0.3159) \\ &= 0.6318 \end{aligned}$$

- (ii) $P(\text{more than } 98\% \text{ of the couriers will be delivered by the next day})$

$$\begin{aligned} P(p > 0.98) &= P\left(Z > \frac{0.98-0.95}{0.011}\right) \\ &= P(Z > 2.7) \end{aligned}$$



$$\begin{aligned}
 &= 0.5 - P(0 < Z < 2.7) \\
 &= 0.5 - 0.4965 \\
 &= 0.0035
 \end{aligned}$$

Example 6.16: It is estimated that 5% of the credit card statements processed in a certain bank contain at least one error. A simple random sample of 500 credit card statements are examined. Find the probability that the proportion of card statements containing at least one error is between 0.04 & 0.075?

Solution:

$$P = 0.05$$

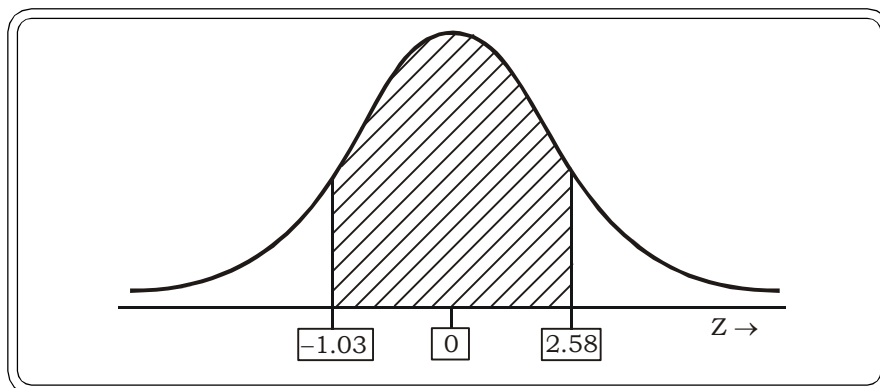
$$N = 500$$

The sample proportion will follow a normal distribution with mean = 0.05 and

$$\begin{aligned}
 \text{Standard Deviation} &= \sqrt{\frac{(0.05)(0.95)}{500}} \\
 &= \sqrt{0.000095} \\
 &= 0.0097
 \end{aligned}$$

Thus,

$$\begin{aligned}
 &P(0.04 < p < 0.07) \\
 &= P\left(\frac{0.04 - 0.05}{0.0097} < Z < \frac{0.075 - 0.05}{0.0097}\right) \\
 &= P(-1.03 < Z < 2.58)
 \end{aligned}$$



$$\begin{aligned}
 &= P(0 < Z < 1.03) + P(0 < Z < 2.58) \\
 &= 0.3485 + 0.4951 \\
 &= 0.8436
 \end{aligned}$$

6.4.4 Student's t – Statistic and its Distribution

We now investigate the distribution of a statistic t defined by

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

where, \bar{x} is the mean of a random sample of n observations

x_1, x_2, \dots, x_n .

μ = population mean

n = sample size and

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

We briefly trace the genesis of this statistic. The central limit theorem states that the distribution of \bar{x} is normal with mean μ and variance $\frac{\sigma^2}{n}$. By converting to the z scale

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow N(0, 1)$$

However when σ is not known we use a unbiased estimator (discussed in chapter 7) of σ viz. s defined as

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\text{i.e. } t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

This will provide a good approximation to the standard normal distribution as long as n is large. However if n is small the distribution of t will show a marked departure from the normal distribution. This fact was first discovered by William Gosset, an employee of the Guinness Brewery. Due to restriction of his employer in publication of his works, he published his research work under the pseudonym 'student'.

And the statistic

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

is known as the student's t - statistic. It follows a student's t - distribution with $(n-1)$ degrees of freedom.

An interesting feature of ' t ' defined above is that both the statistic and its sampling distribution are independent of σ , the population standard deviation.

The student's t statistic following a t -distribution with $(n - 1)$ degrees of freedom is denoted by $t \sim t (n - 1)$

Important properties of the student's t - distribution

- (1) It is a continuous distribution, ranging from $-\infty$ to ∞ .
- (2) It is symmetric about its mean i.e. zero.
- (3) The distribution is generally bell shaped, but is flatter than the normal distribution, with thicker tails. For large samples (say $n \geq 30$), the differences are negligible.
- (4) The shape of the t -distribution depends on a parameter called $v = (n-1)$, which is called its degrees of freedom (df).
- (5) t being a continuous distribution, its probabilities are represented as areas under the curve.

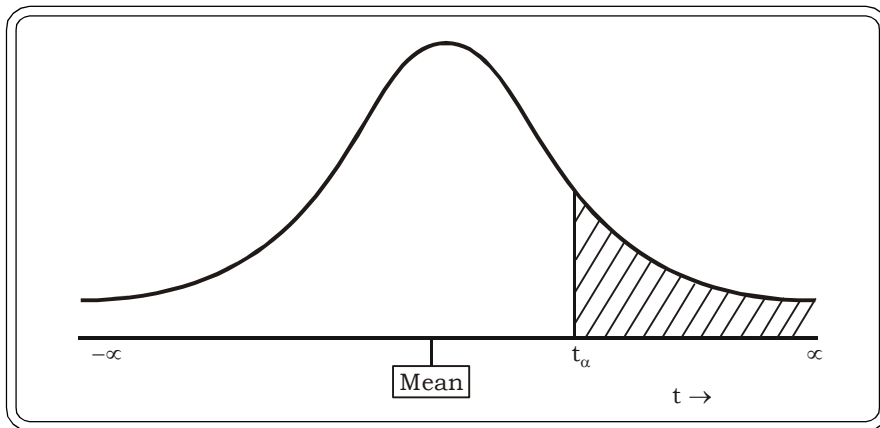


Figure 6.8

Graph of a t - distribution

Table 6.6
Example of a t -table

d.f.	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$
6	1.440	1.943	2.447	3.143
7	1.415	1.895	2.365	2.998
8	1.397	1.860	2.306	2.896

Interpretation of t-value

Consider the value 1.895 under $t_{0.050}$. The value 1.895 represents the ordinate, the area to the right of which is 0.050 (the subscript with t) corresponding to 7 d.f.

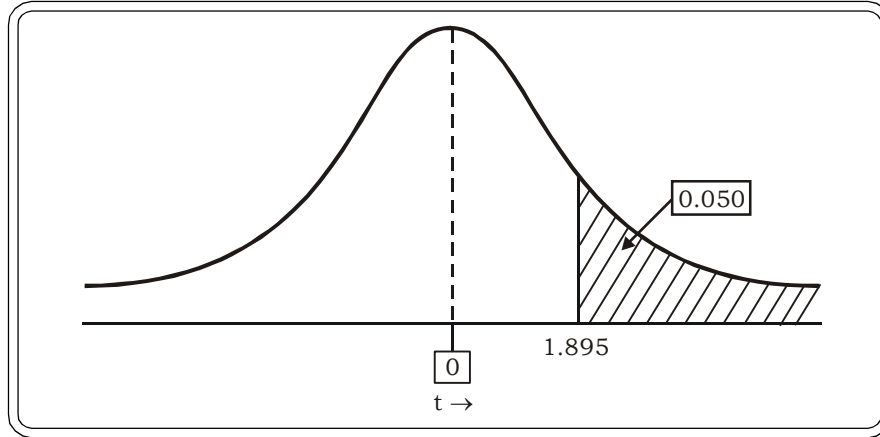


Figure 6.9

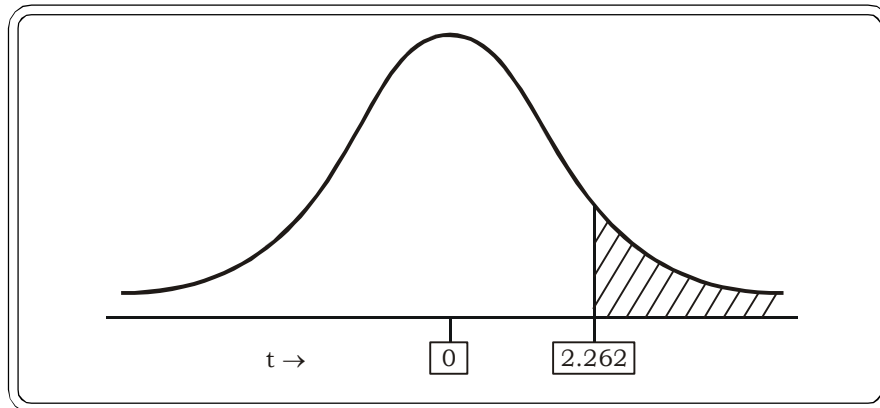
Interpretation of t-value

Example 6.17: A random sample of size 10 is drawn from the normal distribution. Use t-table to find

- (i) $P(t > 2.262)$
- (ii) $P(-1.383 < t < 2.262)$

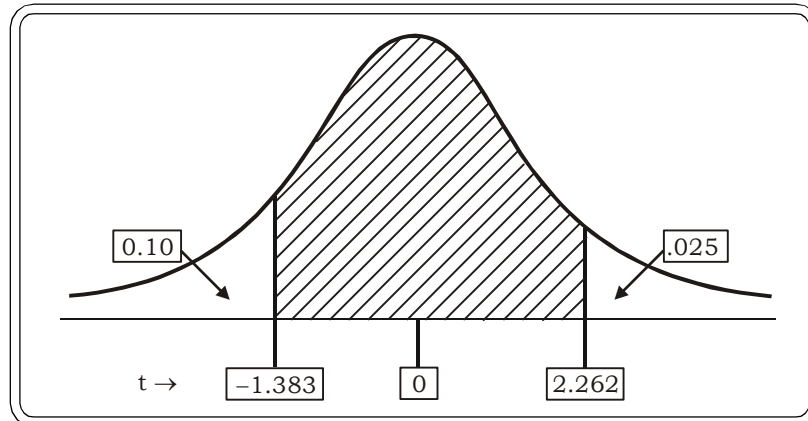
Solution:

- (i) $P(t > 2.262)$



We require the area to the right of 2.262 for degrees of freedom $v = 10 - 1 = 9$. The area from the t-table is 0.025.

(iii) $P(-1.383 < t < 2.262)$



Since the curve is symmetric, the area to the right of 2.262 is 0.025 and the area to the left of -1.383 is 0.100 .

$$\begin{aligned} \text{So } P(-1.383 < t < 2.262) &= 1 - (0.025 + 0.100) \\ &= 1 - 0.125 = 0.875 \end{aligned}$$

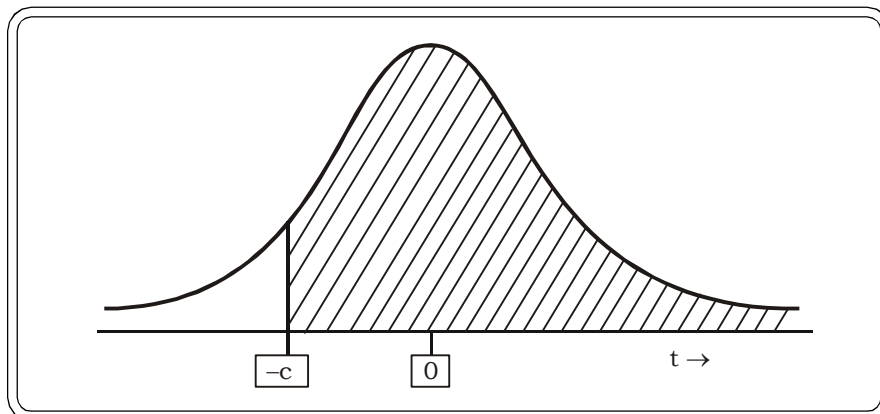
Example 6.18: If T has the t - distribution with 14 degrees of freedom, find c such that

(i) $P(T > -C) = 0.975$

(ii) $P(-C < T < C) = 0.8$

Solution:

(i)



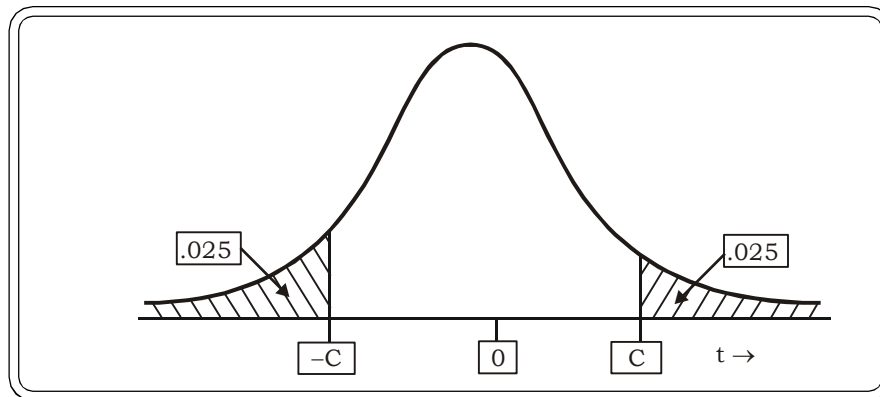
Given

$$P(T > -C) = 0.975$$

$$\Rightarrow 1 - P(T < -C) = 0.975$$

$$P(T < -C) = 0.025$$

$$P(T > -C) = 0.025$$

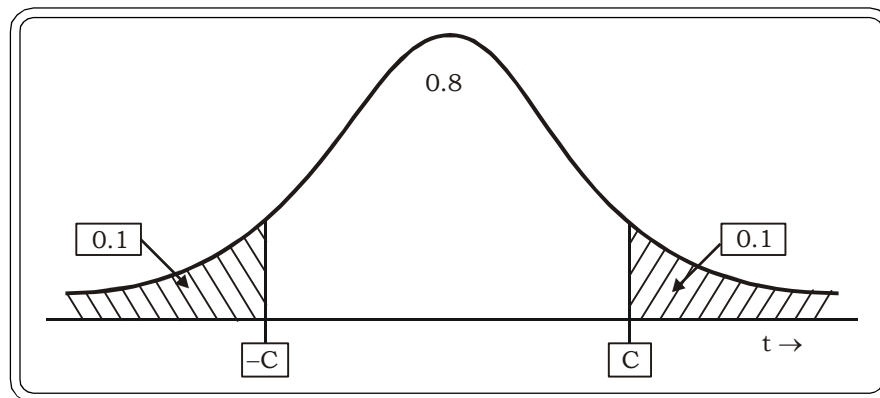


From tables $C = 2.145$, for 14 d.f

Thus $P(T > 2.145) = 0.025$

And $P(T > -2.145) = 0.975$

(ii) $P(-C < T < C) = 0.8$



$$P(T > C) = \frac{1 - P(-C < T < C)}{2} = \frac{1 - 0.8}{2} = 0.1$$

For 14 degrees of freedom the value of C is 1.345. (from the t tables)

Thus $P(-1.345 < T < 1.345) = 0.8$

6.4.5 The Chi - Square Statistic and its Distribution

We next study the distribution of a statistic called the chi - square statistic.

Suppose we have n independent observations from a population, which is normally distributed with mean μ and variance σ^2 . Consider a sample of size n from this population.

Sample values: $x_1 \quad x_2 \quad x_3 \quad \dots \quad x_n$

Each of these values can be treated as an independent random variable with mean μ and variance σ^2 .

Thus

$$E(x_i) = \mu$$

$$V(x_i) = \sigma^2 \quad i = 1, 2, \dots, n$$

Standardizing these values, we get $Z_i = \frac{x_i - \mu}{\sigma}$, $i = 1, 2, \dots, n$

Such that

$$Z_i \rightarrow N(0,1)$$

Consider the sample statistic U defined below, which is the sum of squares of n independent standard normal variables.

$$\begin{aligned} U &= Z_1^2 + Z_2^2 + \dots + Z_n^2 \\ &= \sum_{i=1}^n Z_i^2 \\ &= \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2 \end{aligned}$$

Being the sum of squares of independent random variables U will also be a random variable. This statistic is called a χ^2 square statistic or a chi – square variable. The probability distribution of this Statistic is called the chi – square distribution with n d.f. Symbolically,

$$U = \chi^2 \sim \chi^2(n)$$

This statistic has many applications, some of which are described in chapter 10.

The chi – square statistic does not make any assumptions regarding the population from which the samples are drawn. Hence it is often referred to as a non – parametric test.

Important properties of the chi-square distribution

- (1) The chi – square distribution is a continuous distribution, ranging from 0 to ∞
- (2) The χ^2 values cannot be negative and hence the curve is always on the 1st quadrant.
- (3) The shape of the curve depends on a parameter n, which is called the number of degrees of freedom.

For smaller degrees of freedom the curve is skewed to the right.

As the degrees of freedom increases, the skeweness disappears and for large n ($n \geq 30$) the distribution can be approximated by a normal distribution, as shown in the following figure. 6.10.

Generally, when $n \geq 30$, Fisher's approximation is used. Thus, $\sqrt{2\chi^2}$ is used instead of χ^2 such that $\sqrt{2\chi^2} \longrightarrow N(\sqrt{2n-1}, 1)$

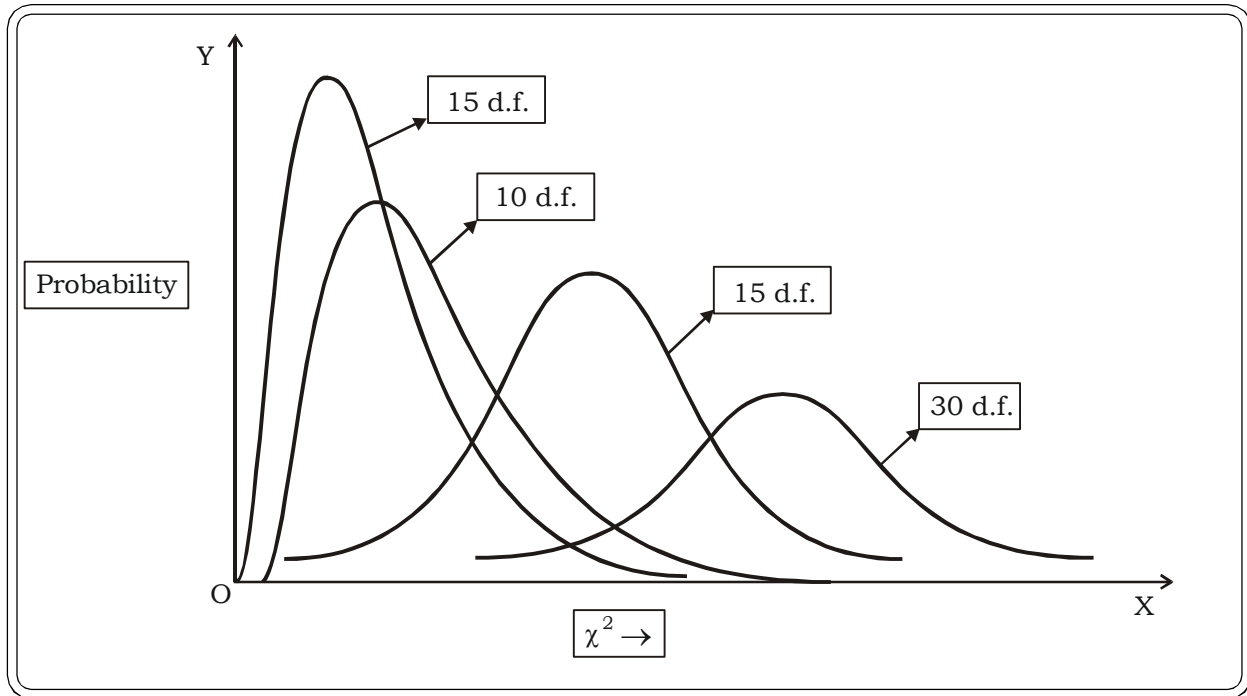


Figure 6.10

Chi - Square Graphs for Different Degrees of Freedom

- (4) Since the distribution is continuous, probabilities are represented as areas under the curve. Consider the following portion of the chi - square table, as an example:

Table 6.7
Example of a χ^2 - Table

Degrees of freedom (ν)	α						
	0.995	0.99	0.975	0.95	0.05	0.025	0.01
	$\chi^2_{0.995}$	$\chi^2_{0.99}$	$\chi^2_{0.975}$	$\chi^2_{0.95}$	$\chi^2_{0.05}$	$\chi^2_{0.025}$	$\chi^2_{0.01}$
1	0.676	0.872	1.237	1.635	12.592	14.449	16.812
2	0.989	1.239	1.690	2.167	14.067	16.013	18.475
3	1.344	1.646	2.180	2.733	15.507	17.535	20.090
4	1.735	2.088	2.700	3.325	16.919	19.023	21.666

Interpretation of χ^2 -values

Let, $\chi^2(4, 0.995) = 1.735$

For degrees of freedom 4, the area to the right of the curve from the point 1.735 is 0.995

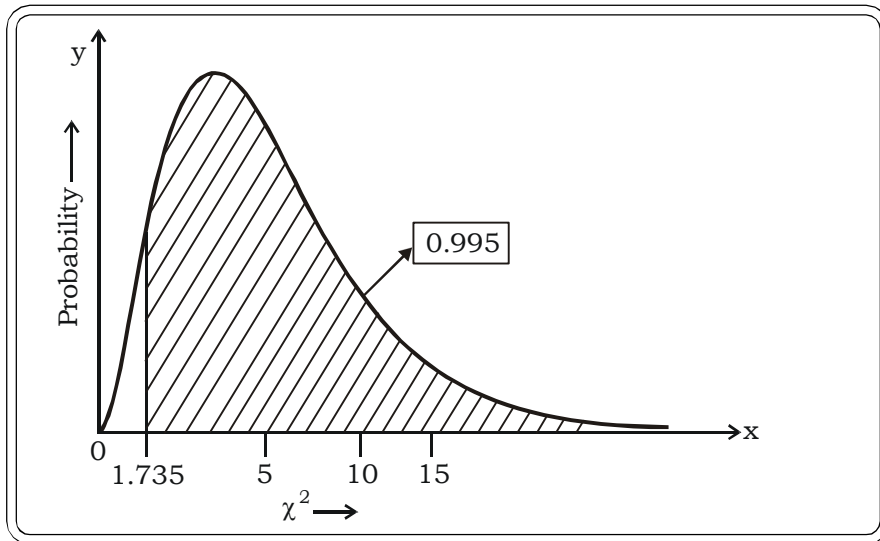


Figure 6.11
Interpretation of χ^2 - value

Example 6.19: A random sample of size 6 is drawn from a normal distribution. Use the chi – square tables to find

(i) $P(\chi^2 > 12.592)$

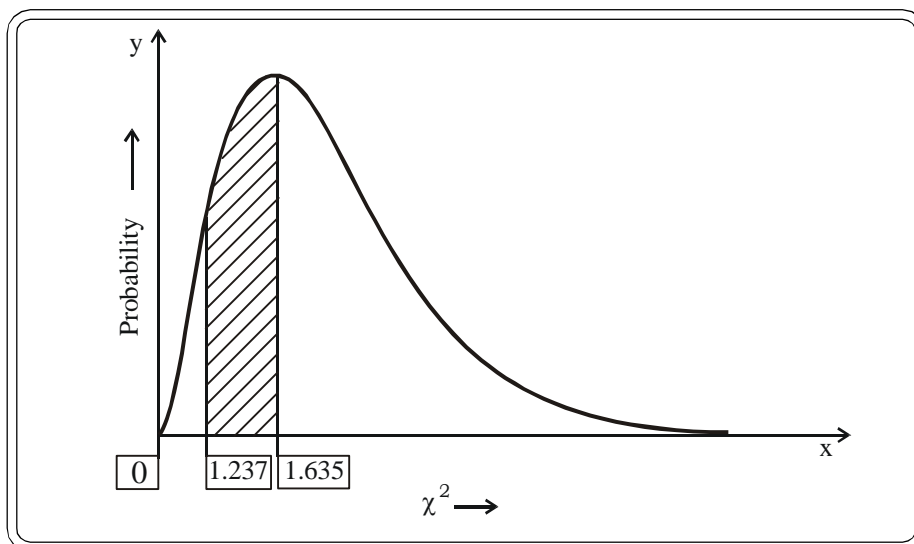
(ii) $P(1.237 < \chi^2 < 1.635)$

Solution:

(i) $P(\chi^2 > 12.592)$ will be the area of the graph to the right of 12.592 for 6 degrees of freedom. From the table, this value is .05

Thus $P(\chi^2 > 12.592) = 0.05$

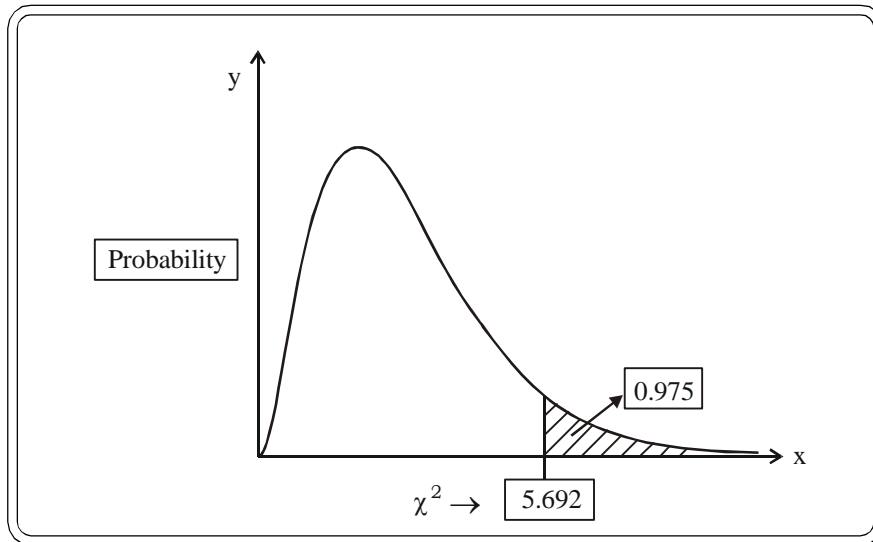
(ii) $P(1.237 < \chi^2 < 1.635)$



$$\begin{aligned}
 &= (\text{Area to the right of } 1.237) - (\text{Area to the right of } 1.635) \\
 &= 0.975 - 0.95 \\
 &= 0.025
 \end{aligned}$$

Example 6.20 If u has a chi – square with 14 degrees of freedom. Find c such that $P(u > c) = 0.975$

Solution:



From the chi square table area to the right of 5.692 is 0.975

$$C = 5.692$$

$$\text{Thus } P(u > 5.692) = 0.975$$

6.4.6 The F – Statistic and its Distribution

Let χ_1^2 be a chi – square variable with n_1 degrees of freedom.

Let χ_2^2 be a chi – square variable with n_2 degrees of freedom.

Let χ_1^2 and χ_2^2 be independent.

Then the F statistic is defined by the ratio:

$$F = \frac{\frac{\chi_1^2}{n_1}}{\frac{\chi_2^2}{n_2}}$$

The statistic F is said to follow a F distribution with n_1 and n_2 degrees of freedom. The order of the degrees of freedom is important. The degrees of freedom associated with the chi – square in the numerator should precede the degrees of freedom associated with the denominator. In general, it is denoted by

$$F \sim F(n_1, n_2)$$

Important properties of the F – Distribution

- (1) Since F – distribution depends only on 2 parameters n_1 and n_2 they are a part of the non parametric family of distributions.

- (2) It is the ratio of two χ^2 divided by their degrees of freedom, it can assume only positive values. Range of the distribution is from 0 to ∞ .
- (3) The graph of the F – distribution is skewed to the right but tends to symmetry as the number of degrees of freedom increases.
- (4) The F – distribution with n_1 and n_2 degrees of freedom and leaving an area of α in the right tail is usually denoted by $F(n_1, n_2, \alpha)$.

General levels of α are $\alpha = 0.1, 0.05, 0.025, 0.01$.

Consider the following example of the F – table, for $\alpha = 0.1$

Table 6.8
Example of F-Table

$v_2 \backslash v_1$	1	2	3	4	5
5	4.06	3.78	3.62	3.52	3.45
6	3.78	3.46	3.29	3.18	3.11
7	3.59	3.26	3.07	2.96	2.88

Interpretation of F-values

Let $F(4, 5, 0.1) = 3.52$.

This means for 4 and 5 degrees of freedom and $\alpha = 0.1$, the area to the right of 3.52 is 0.10

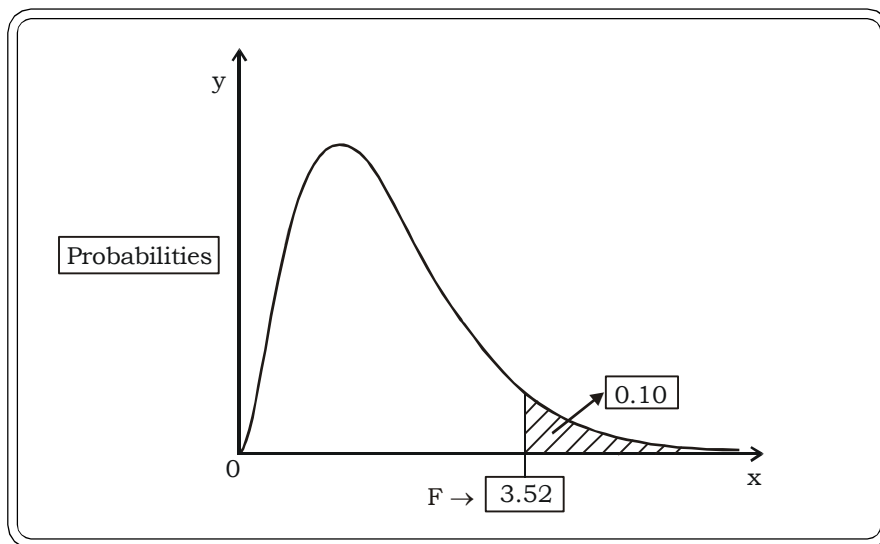


Figure 6.12

Interpretation of F-values

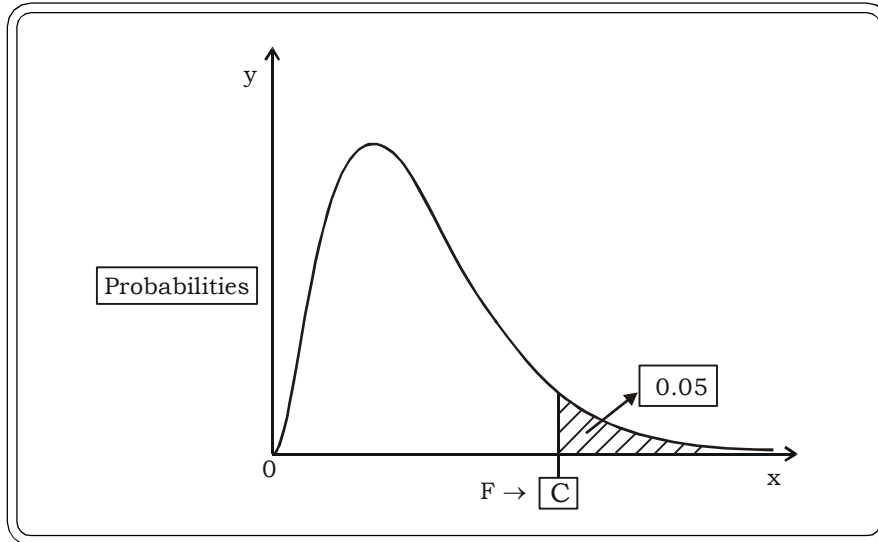
Example 6.21: Suppose F has F – distribution with $n_1 = 15$ and $n_2 = 9$ degrees of freedom. Find c such that

(i) $P(F > c) = 0.05$

(ii) $P(F > c) = 0.025$

Solution:

(i)



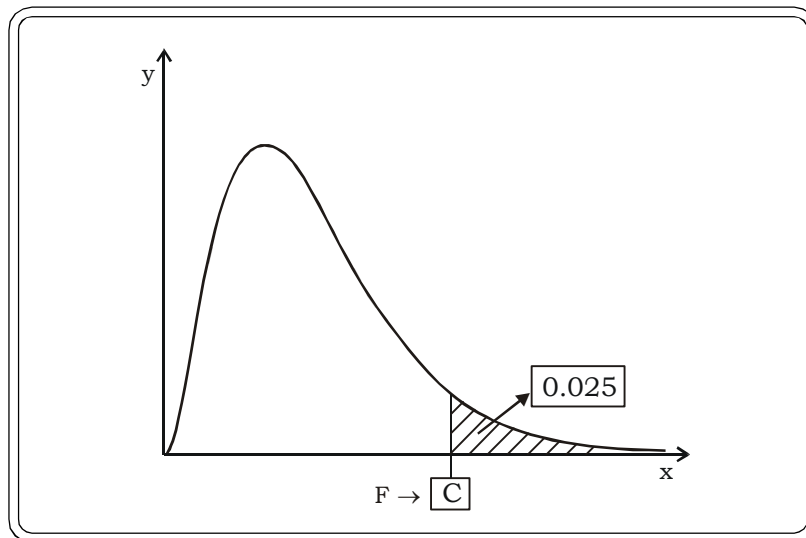
We select the F – table with $\alpha = 0.05$.

From this table, the value corresponding to $n_1 = 15$ and $n_2 = 9$ is 3.01

Thus $c = 3.01$

Thus $P = (F > 3.01) = 0.05$

(iii) In this case, we choose the F table with $\alpha = 0.025$.



The area to the right of c is 0.025. Corresponding to degrees of freedom 15 and 9 this value is 3.77.

Thus $c = 3.77$

And $P(F > 3.77) = 0.025$.

Table 6.9
Summary of Sample Statistic and their Distribution

Sample statistic	Distribution
\bar{x} - The sample mean	$N\left(\mu, \frac{\sigma^2}{\sqrt{n}}\right)$
p - The sample proportion	$N\left(P, \frac{P(1-P)}{n}\right)$
t-statistic: $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$	t - distribution with $(n - 1)$ d.f.
χ^2 statistic: $\chi^2 = \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma}\right)^2$	χ^2 - distribution with n d.f.
F-statistic: $F = \frac{\chi_1^2/n_1}{\chi_2^2/n_2}$	F - distribution with (n_1, n_2) d.f.

6.5 EXERCISES

- 6.1 The 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 we pick up a random sample of size 5, what is the probability that the sample mean will be between 9.95mm and 10.05 mm? **(MBA, DU, 1997)**
- 6.2 A manufacturer of watches has determined from previous experience that 3% 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? **(MBA, Delhi Univ., 2000)**
- 6.3 Car Stereos of a manufacturer 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 random samples of 125 stereos of each manufacturer are tested, what is the probability that the manufacturer A stereos will have a mean lifetime which is at least (i) 160 hours more than the, manufacturer B stereos and (ii) 250 hours more than the manufacturer B stereos? **(MBA, Delhi Univ., 1999)**
- 6.4 The average annual starting salary for MBA (Marketing majors) is Rs.3, 40,000. Assume that for the population of MBA (Marketing majors), the average annual starting salary is $\mu = 3,40,500$ 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 Rs.2, 500 of the population mean for each of the sample sizes: 50, 100 and 200? What is your conclusion? **(MBA, Delhi Univ., Oct., 2003)**

- 6.5 What is sampling? Explain the importance of sampling in solving business problems. Critically examine the well-known methods of probability sampling and non- probability sampling?
(MBA, DU, 1998)
- 6.6 Answer the following questions:
- Explain any four sampling methods you are aware of.
 - What are the various types of sampling?
 - Differentiate between 'sample' and 'population'. Point out their advantages and limitations.
- 6.7 What are the advantages of sampling? **(MBA, Madurai-Kamaraj, 2001)**
- 6.8 A commercial bank is deciding whether to open a branch in a new community or not. They have decided on the following sampling rule. Take a random sample of 100 families, if their average income is \$22, 000 or more, then they will open the branch, otherwise they will not
- What is the probability that the bank will open a branch in a new community where the annual family income is \$21,500 with a standard deviation of \$1500?
(MBA, Bharathidasan Univ., 2003)
 - What is the probability that the bank will not open a branch in a community with an average family income of \$22,500 with a standard deviation of \$1500?
(MBA, Bharathidasan Univ., 2001)
- 6.9 The time between two arrivals of customers at the bank is normally distributed with a mean of 5 times and a standard deviation of 1 minute. If random sample of 30 such times between successive arrivals is taken, what is the probability that the sample mean will be less than 2 minutes?
(MBA, DU, 2001)
- 6.10 A simple random sample of 400 batteries is taken out of a batch of 1,000 for testing as to the average life of these batteries in the sample in hours of use. The batteries are known to have an average life of 120 hours with a standard deviation of 16 hours. The battery life is normally distributed. What is the probability that:
- The average life of batteries would turn out to be less than 121 hours.
 - The average life of these batteries would be between 120 hours and 121 hours.
- 6.11 A hotel chain employs over 300 people. The workers' ages are approximately normally distributed with an average of 39 and a standard deviation of 5.4 years. The company is thinking of buying medical insurance for the workers. The insurance company wants to take a random sample of 25 workers to determine the average age before quoting a price. This sample was taken randomly from the master list of all employees.
- What is the probability that the sample average will be less than 30 years?
 - What is the value of the standard error of the mean associated with this sample?
 - What can the insurance company do to reduce the standard error of the mean?
- 6.12 A music production company produces classical music on CDs. Previous experience has shown that 20% of all such CDs produced do not sell enough to cover the cost of production and marketing and hence are considered failures. If the company adds 60 new titles to its list in a given year, what is the probability that:
- There will be 10 or more failures.
 - There will be between 15% and 20% failures.
 - There will be less than 6 failures.

- 6.13 Suppose 5% of the tubes produced by a machine are defective. If a sample of 100 tubes is inspected at random,
- Find the expected proportion of defectives in the sample.
 - Find the variance of the proportion of defectives in the sample.
 - Find the approximate distribution of the sample proportion.
 - Find the probability that the proportion of defectives will exceed 0.15
- 6.14 If 60% of the population feels that the Indian Prime Minister is doing a satisfactory job, find the approximate probability that in a sample of 900 people interviewed at random, the proportion who share this view will
- Exceed 0.65
 - Be less than 0.56
- 6.15 It is hypothesized that the proportion of individuals in the population with blood type O is 0.3. If this hypothesis is correct, what is the probability that in a random sample of 400, the proportion of blood type O individuals will be less than 0.25 or greater than 0.35?
- 6.16 Suppose it is known that 5% of forms processed by a clerical pool contain at least one error. If a simple random sample of 475 forms is examined, what is the probability that the proportion containing at least one error will be between 0.03 and 0.075?
- 6.17 Define the chi – square statistic. State important properties of the chi – square distribution.
- 6.18 Define the t – statistic and state important properties of the t – distribution.
- 6.19 Define the F – statistic. Mention some of the important properties of the F – distribution.
- 6.20 If T has a t – distribution with 10 degrees of freedom, find c such that
- $P(T > C) = 0.975$
 - $P(T > C) = 0.90$
 - $P(T > C) = 0.99$
- 6.21 Find the following probabilities if T has t-distribution with given degrees of freedom.
- $P(T > 2.179)$ with 12 degrees of freedom
 - $P(T < - 2.821)$ with 9 degrees of freedom
 - $P(-1.345 < T < 2.624)$ with 14 degrees of freedom
- 6.22 Let x^2 has a χ^2 distribution with the given degrees of freedom, find
- $P(x^2 > 4.404)$ with 12 degrees of freedom
 - $P(x^2 < 0.484)$ with 4 degrees of freedom
 - $P(0.831 < x^2 < 11.070)$ with 5 degrees of freedom
- 6.23 Consider a F – distribution with 15 and 9 degrees of freedom. Find C such that
- $P(F < C) = 0.90$
 - $P(F > C) = 0.05$



7

Theory of Estimation and Testing of Hypothesis



Structure

- 7.1 Introduction
- 7.2 Estimation
 - 7.2.1 Point Estimation
 - 7.2.1.1 Point Estimator of Population Mean
 - 7.2.1.2 Point Estimator of Population Proportion
 - 7.2.1.3 Point Estimator of Population Variance
 - 7.2.2 Interval Estimation
 - 7.2.2.1 Interval Estimator of Population Mean
 - 7.2.2.2 Interval Estimator of Difference of Two Means
 - 7.2.2.3 Interval Estimator of Single Population Proportion
 - 7.2.2.4 Interval Estimator of Difference of Two Proportions
 - 7.2.2.5 Determination of Sample Size
- 7.3 Testing of Hypothesis
 - 7.3.1 Null and Alternative Hypothesis
 - 7.3.2 Type I Error Type II Error
 - 7.3.3 One-Tailed Test Two-Tailed Test
 - 7.3.4 One Sample Tests
 - 7.3.4.1 One Sample Z Test for Mean
 - 7.3.4.2 One Sample t Test for Mean
 - 7.3.4.3 One Sample Z Test for Proportion
 - 7.3.5 Two Sample Tests
 - 7.3.5.1 Two Sample Z Test for Difference of Two Means
 - 7.3.5.2 Two Sample t Test for Difference of Two Means
 - 7.3.5.3 Paired t Test (for correlated samples)
 - 7.3.5.4 Two Sample Z Test for Difference of Two Proportions
- 7.4 Caselets
- 7.5 Excel Guide
- 7.6 Exercises

7.1 INTRODUCTION

A population is a well-defined group of subjects: e.g., individuals, firms, countries, cities, etc. Inferential statistics are used to draw inferences about a population from a sample. It involves learning something about a population, given the availability of a sample from that population. By “learning something” we mean to obtain approximations or estimates of some parameters that characterize the population: mean, variances, correlations etc.

For example, consider an experiment in which 10 subjects who performed a task after 24 hours of sleep deprivation scored 12 points lower than 10 subjects who performed after a normal night’s sleep. Is the difference real or could it be due to chance? How much larger could the real difference be than the 12 points found in the sample? These are the types of questions answered by inferential statistics. There are two main areas in inferential statistics:

- (i) estimation and
- (ii) hypothesis testing.

This chapter describes both these areas of inferential statistics along with applications.

7.2 ESTIMATION

Estimation theory is a branch of statistics that deals with estimating the values of unknown parameters based on measured/empirical data. The parameters describe the physical scenario or object that answers a question posed by the estimator. For example, it is desired to estimate the proportion of a population of voters who will vote for a particular candidate. That proportion is the unknown parameter; an estimate of this unknown parameter is based on a small random sample of voters.

The entire purpose of estimation theory is to arrive at an estimator, and preferably an implementable one that could actually be used. The estimator takes the measured data as input and produces an estimate of the parameters. It is also preferable to derive an estimator that exhibits optimality. An optimal estimator would indicate that all available information in the measured data has been extracted, for, if there was unused information in the data then the estimator would not be optimal.

Given the distribution of a variable in a population, we obtain the results about the distributions of various quantities, such as the mean and variance, calculated from sample observations. Such a quantity is called a statistic. These results are of direct interest in the planning of sampling enquires, as they enable the investigator to estimate the precision attainable with a sample of a given size, and hence help him to decide how large a sample should be taken.

Estimator and Estimate

When the sample has been taken, what sort of inferences can be drawn about the population, on the basis of the sample? We don’t know the characteristics of the population. We have taken one random sample and wish to use our knowledge of sampling theory to make whatever inference can be made about the population. One fundamental difficulty usually arises. The expressions of sampling variation given by the various formulae for standard errors or variances usually involve some

parameters of the population. For instance, the standard error of the sample mean is $\frac{\sigma}{\sqrt{n}}$. If we are attempting to make an inference about a normal distribution on the basis of one random sample,

we may know the sample size, n , but not the population standard deviation. We cannot, therefore, calculate the standard error exactly.

An **estimate** is an educated guess about an unknown quantity or outcome based on known information. A rule that tells how to calculate an estimate based on the measurements contained in a sample is called an **estimator**. Thus, an estimator is a sample quantity i.e. a statistic used to estimate a population quantity and a specific observed numerical value from a particular sample is an estimate. For example, the “sample mean” \bar{x} is an estimator for the population mean μ .

And, for a random sample of size 5, viz.

5, 6, 8, 5, 6

the sample mean

$$\bar{x} = \frac{5+6+8+5+6}{5} = 6$$

This particular value of the mean i.e. 6 is an estimate.

Thus, the estimator is a random variable and an estimate is a computed value from a given sample.

Characteristics of Good Estimators

A good estimator should satisfy the following characteristics:

- (i) Unbiasedness
- (ii) Consistency
- (iii) Efficiency
- (iv) Sufficiency

Unbiasedness

A statistic is biased if, in the long run or when we take repeated samples of the same size, it consistently over or underestimates the parameter it is estimating. For estimation, an estimator “on the average” should give the parameter it is supposed to estimate. Thus, technically it is biased if its expected value is not equal to the parameter. A stop watch that is a bit fast gives biased estimates of elapsed time. Bias in this sense is different from the notion of a biased sample. A statistic is positively biased if it tends to overestimate the parameter; a statistic is negatively biased if it tends to underestimate the parameter. An unbiased statistic is not necessarily an accurate statistic. In statistical term the expected value of the sample estimate is considered to be an unbiased estimator if it equals the population parameter.

Thus, an estimator $\hat{\theta}$ of a parameter θ is said to be an unbiased estimator if

$$E(\hat{\theta}) = \theta$$

For example, the sample mean \bar{x} , is an unbiased estimate of the population mean μ which means that the expected value of the sample mean is equal to the population mean μ . Symbolically,

$$E(\bar{x}) = \mu$$

Consistency

An estimator is consistent if the estimator tends to get closer to the parameter it is estimating as the sample size increases. The sample mean, as an estimator of the population mean is consistent i.e.

$$\bar{x} \longrightarrow \mu \text{ as } n \longrightarrow \infty$$

Thus, if n is very large, the probability that \bar{x} is close to μ will be almost close to 1. A consistent estimator must be at least asymptotically unbiased. An estimate is said to be asymptotically unbiased if the bias tends to zero for a large number of observations.

Efficiency

The efficiency of a statistic is the degree to which the statistic is stable from sample to sample. That is, the less subject to sampling fluctuation a statistic is, the more efficient it is. For example, if we have two unbiased estimators θ_1 and θ_2 of a parameter θ , the one with the smaller variance is said to be the more efficient estimator.

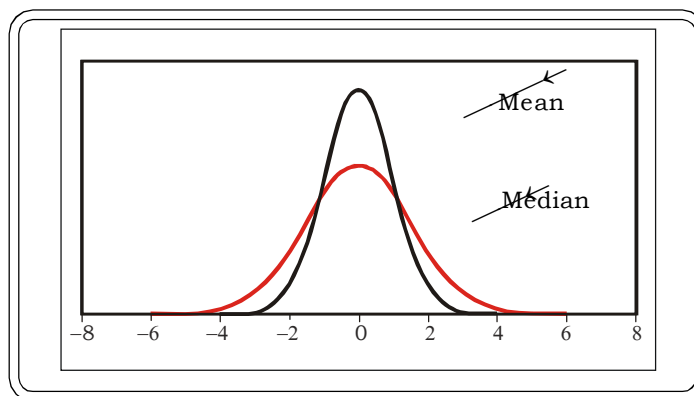


Figure 7.1

Sampling Distributions for Mean

For normal populations, the sample mean is the most efficient estimator of μ .

Thus, the efficiency of a statistic is measured relative to the efficiency of other statistics and is therefore often called the relative efficiency. If statistic A has a smaller standard error than statistic B, then statistic A is more efficient than statistic B. In figure 7.1, the mean has a smaller variance than the median and hence is more efficient.

The relative efficiency of two statistics may depend on the distribution involved. For instance, the mean is more efficient than the median for normal distributions but not for some extremely skewed distributions.

Sufficiency

An estimator is said to be a sufficient estimator if it considers all the information about the population parameter present in the sample for the purpose of estimating the parameter. For example, mean uses all the sample values in its computation while mode and median do not. And so mean is the better estimator than the other two measures of average in terms of sufficiency.

A particular estimator may or may not satisfy some or all of the four criteria. In general the sample mean is a good estimator of the population mean and satisfies all the four criteria i.e. it is unbiased, consistent, efficient and sufficient.

Also, for estimating the population variance σ^2 , the statistic

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

based on a sample x_1, x_2, \dots, x_n of n observations, is the best estimator. Also, the sample proportion is a good estimator of the population proportion.

Types of Estimators

There are two types of estimators that are commonly used viz.,

- (i) Point estimator and
- (ii) Interval estimator

7.2.1 Point Estimation

When a parameter is being estimated, the estimate can be either a single number or it can be a range of values. When the estimate is a single number, the estimate is called a point estimate. For instance, while planning a trip from Delhi to Agra we might estimate the distance as x kms, the mileage as y kms / litre and the price of petrol as Rs z /litre. This information can now be put together to estimate the cost of the entire trip which can be viewed as a point estimate.

As another example of a point estimate, assume that we want to estimate the mean time it takes for 12-year-olds to run 100 yards. The mean running time of a random sample of 12-year-olds would be an estimate of the mean running time for all 12-year-olds. Thus, the sample mean say M , would be a point estimate of the population mean, μ .

Often, point estimates are used as parts of other statistical calculations. For example, a point estimate of the standard deviation is used in the calculation of a confidence interval for μ . Point estimates of parameters are often used in the formulae for significance testing. These will be dealt with in the later sections.

The advantage of point estimation is that it yields a precise value. The disadvantage is that the confidence that the value selected is correct is low. Point estimation is rare. Point estimates are not usually as informative as confidence intervals. Their importance lies in the fact that many statistical formulae are based on them.

7.2.1.1 Point Estimator of Population Mean

The best point estimator of the *population mean* μ is the sample mean \bar{x} .

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

It is unbiased, consistent, sufficient and most efficient point estimator.

7.2.1.2 Point Estimator of Population Proportion

Let x denote the number of units in a sample of size n possessing a certain attribute.

Then, the sample proportion,

$$p = \frac{x}{n}$$

is an unbiased estimator of the population proportion, and also the best estimator of the population proportion.

7.2.1.3 Point Estimator of Population Variance

The formula for the variance computed in the population, σ^2 is different from the formula for an unbiased estimator of variance i.e. s^2 , computed from sample information. The two formulae are shown below.

Let X_1, X_2, \dots, X_N be the N units of a population

$$\text{Population variance: } \sigma^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}$$

Let x_1, x_2, \dots, x_n be a sample of n units from this population.

$$\text{Unbiased estimator of population variance: } \hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

If the formula for sample variance is used as

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

it will give a biased estimate of population variance (σ^2). However, by far the most common formula for computing variance in a sample is

$$\hat{\sigma}^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

which gives an *unbiased estimate* of σ^2 . Since samples are usually used to estimate parameters s^2 is the most commonly used measure of variance.

The difference between the two formulae (formula of population variance and sample variance) is that the denominator is n for σ^2 and is $n-1$ for s^2 . That there should be a difference in formulae is very counterintuitive. To understand the reason that $n-1$ rather than n is needed in the denominator of the formula for s^2 , consider the problem of estimating σ^2 when the population mean, μ , is already known.

Suppose that we knew that the mean amount of practice it takes for student pilots to master a particular manoeuvre is 12 hours. If we sampled one pilot and found he or she took 14 hours to master the manoeuvre, what would be our estimate of σ^2 ? The answer lies in considering the definition of variance: It is the average squared deviation of individual scores from μ .

With only one score, we have one squared deviation of a score from μ . In this example, the one squared deviation is: $(x - \mu)^2 = (14-12)^2 = 4$. This single squared deviation from the mean is the best estimate of the average squared deviation and is an unbiased estimate of σ^2 . Since it is based on only one score, the estimate is not a very good estimate although it is still unbiased. It follows that if μ is known and n scores are sampled from the population, then an unbiased estimate of σ^2 could

be computed with the following formula: $\sum (x - \mu)^2 / n$. Now it is time to consider what happens when μ is not known and \bar{x} is used as an estimate of μ . Which value is going to be larger for a sample of n values of x : $\sum (x - \bar{x})^2 / n$ or $\sum (x - \mu)^2 / n$? Since \bar{x} is the mean of the n values of x and since the sum of squared deviations of a set of numbers from their own mean is smaller than the sum of squared deviations from any other number, the quantity $\sum (x - \bar{x})^2 / n$ will always be smaller than $\sum (x - \mu)^2 / n$.

The argument goes that since $\sum (x - \bar{x})^2 / n$, as an estimate of σ^2 is always smaller than $\sum (x - \mu)^2 / n$ then $\sum (x - \bar{x})^2 / n$, must be biased and will have a tendency to underestimate σ^2 . It turns out that dividing by $n - 1$ rather than by n increases the estimate just enough to eliminate the bias exactly.

Another way to consider about why we divide by $n - 1$ rather than by n has to do with the concept of degrees of freedom. When μ is known, each value of x provides an independent estimate of σ^2 : Each value $(x - \mu)^2$ is an independent estimate of σ^2 . The estimate of σ^2 based on n , x 's is simply the average of these n independent estimates. Since the estimate of σ^2 is the average of these

n estimates, it can be written as: $\frac{\sum (x - \mu)^2}{df}$ where there are n degrees of freedom and therefore $df = n$. When μ is not known and has to be estimated with \bar{x} , the n values of $(x - \bar{x})^2$ are not independent because if we know the value of \bar{x} and the value of $n - 1$ of the x 's, then we can compute the value of the n 'th x exactly.

The number of degrees of freedom an estimate is based upon is equal to the number of independent scores that went into the estimate minus the number of parameters estimated en route to the estimation of the parameter of interest. In this case, there are n independent scores and one parameter (μ) is estimated en route to the estimation of the parameter of interest, σ^2 . Therefore the estimate has $n - 1$ degrees of freedom. The formula for s^2 can then be written as: $\sum (x - \bar{x})^2 / df$ where $df = n - 1$. Naturally, the greater the degrees of freedom, the closer the estimate is likely to be to σ^2 .

Table 7.1
Commonly used Point Estimators

Estimator	Parameter
Sample Mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	Population Mean: μ
Sample Standard Deviation: $s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$	Population Variance: σ
Sample Proportion: $p = \frac{x}{n}$	Population Proportion: P

Example 7.1 The following are the weights of four bags of rice (in kgs.) chosen at random from a lot of 100 bags: 102kgs 100kgs 98kgs 97kgs. Find best estimates of

- (i) The true mean weight of all the bags.
- (ii) The true variance of the weight of all bags.
- (iii) The standard deviation of weights.

Solution:

- (i) The true mean weight of all the bags

$$\begin{aligned}\bar{x} &= \frac{102 + 100 + 98 + 97}{4} = \frac{397}{4} \\ &= 99.25 \text{ kgs.}\end{aligned}$$

- (ii) The estimate of variance is

$$\begin{aligned}s^2 &= \frac{1}{n-1} \sum_{i=1}^4 (x_i - \bar{x})^2 \\ &= \frac{1}{3} [(102 - 99.25)^2 + (100 - 99.25)^2 + (98 - 99.25)^2 + (97 - 99.25)^2] \\ &= \frac{1}{3} [7.56 + 0.56 + 1.56 + 5.06] \\ &= 4.91\end{aligned}$$

- (iii) Estimate of standard deviation = 2.22 kgs.

Example 7.2: In a random sample of 400 individuals, 76 wear contact lenses. Estimate the proportion of people in the population who wear contact lenses.

Solution:

$$n = 400 \text{ (sample size)}$$

$$x = 76 \text{ (Number of people who wear contact lenses)}$$

Estimate of proportion of people who wear contact lenses

$$\begin{aligned}\hat{p} &= \frac{x}{n} = \frac{76}{400} \\ &= 0.19\end{aligned}$$

i.e. 19% of the population may be expected to wear contact lenses.

Example 7.3: Out of 100 compact fluorescent lamps tested in a laboratory, 68 lasted beyond 300 hours. Find a point estimate of the true proportion of compact fluorescent lamps that will last beyond 300 hours.

Solution:

$$n = 100$$

A point estimate of the true proportion of fluorescent lamps that last beyond 300 hours is

$$\frac{68}{100} = 0.68$$

i.e. 68% of the CFL's will last beyond 300 hours.

Example 7.4: A publisher wants to know how many copies of a book needs to be printed for the next year. When 200 students of an institute were interviewed it was found that 170 students had purchased the book. If the total number of students of the particular course in that region is 100, 000, obtain an estimate of the number of books the publisher should print.

Solution:

Number of students surveyed: $n = 200$.

Number of students who purchased the book: $x = 170$.

A good estimate of the proportion of students who buys the book is:

$$p = \frac{170}{200} = 0.85$$

Estimated number of students = 1,00,000

An estimated of the number of books the publisher should print for the next year is:

$$0.85 \times 1,00,000 = 85, 000 \text{ copies}$$

7.2.2 Interval Estimation

When the estimate is a range of scores or values, the estimator is called an interval estimator. In interval estimation, one gives a range of values along with a level of confidence. The confidence intervals are used for interval estimates. The advantage of interval estimation is that the investigator can be fairly confident that the population parameter lies within the confidence interval. The disadvantage is that interval estimation gives a range of values, not a specific value, so it is less precise than point estimation. However, interval estimation is the more common type of estimation.

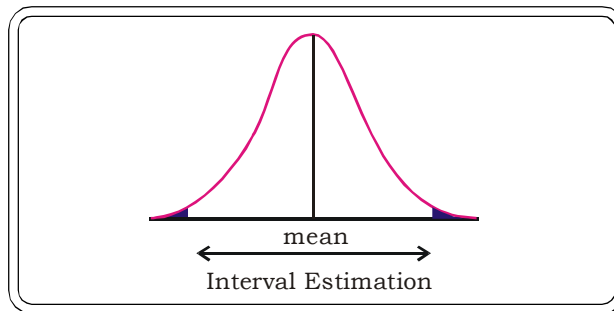


Figure 7.2

Interval Estimation

Confidence Interval

A confidence interval for a parameter is an interval computed from sample data containing the true value of the parameter with a certain level of confidence. For example, if we go back to the Delhi Agra trip example, alternatively, we might also estimate the distance between Delhi and Agra to be between x_1 kms and x_2 kms, the mileage between y_1 and y_2 kms / litre and so on and finally put together this information to arrive at a range of the cost of the trip. For example, it might come out to be between Rs 4000 and Rs 5000. This is an interval estimate.

With a 95% confidence interval for a sample mean, 95% of all samples of the same size will contain the true population mean. Which is very close to saying that the true population mean has a 95% chance of falling within the confidence interval. Another way of putting this is that the interval will include the unknown parameter with probability 0.95.

A confidence interval has the form:

$$\text{estimate} \pm \text{margin of error}$$

Confidence intervals get wider as the confidence increases:

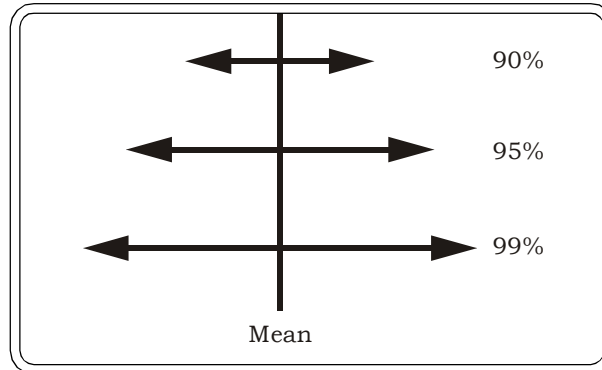


Figure 7.3

Confidence Interval for Different Confidence Levels

Confidence intervals get narrower as sample size increases:

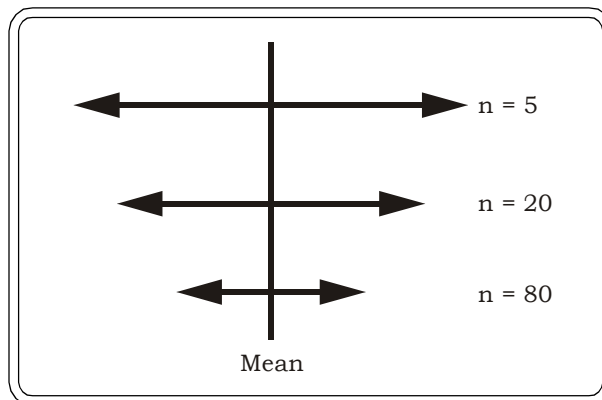


Figure 7.4

Confidence Interval and Sample Sizes

Level of Confidence

The researcher selects a level of confidence to be used in interval estimation. In general, the greater the degree of confidence, the wider the confidence interval must be. Typical confidence levels are 90%, 95%, and 99%. 90% confidence level would mean that 90% of all samples of the same size will contain the true population mean. Confidence levels are related to the alpha level, $\alpha = 1 - \text{confidence level}$. Example $\alpha = 1 - .95 = 0.05$.

90% confidence level ($\alpha = 0.10$)

95% confidence level ($\alpha = 0.05$)

99% confidence level ($\alpha = 0.01$)

7.2.2.1 Interval Estimator of Population Mean

Confidence interval for population mean μ , when population standard deviation is known

Assumptions:

- (1) The parent population follows a normal distribution.
- (2) The population standard deviation σ is known.
- (3) Values or observations are sampled randomly and are independent.

This section explains how to compute a confidence interval for the mean of a normally-distributed variable for which the population standard deviation is known. In practice, the population standard deviation is rarely known. However, learning how to compute a confidence interval when the standard deviation is known is an excellent introduction of how to compute a confidence interval when the standard deviation has to be estimated.

Three quantities are used to compute the confidence interval for μ :

- (i) The sample mean : \bar{x}
- (ii) The standard error of the mean : $\frac{\sigma}{\sqrt{n}} = \sigma_{\bar{x}}$.
- (iii) $Z_{\frac{\alpha}{2}}$: which represents the value on the Z scale such that the area to the right of $Z_{\frac{\alpha}{2}}$ is $\left(\frac{\alpha}{2}\right)$.

By symmetry, the area to the left of $-Z_{\frac{\alpha}{2}}$ is also $\left(\frac{\alpha}{2}\right)$. The remaining area $(1 - \alpha)$ is at the centre. This is shown in figure 7.5 below:

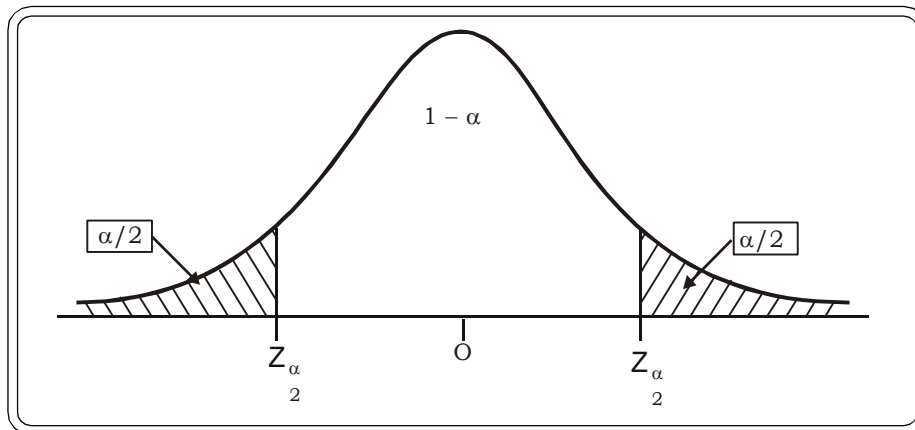


Fig. 7.5

Critical value for $(1 - \alpha)$ 100% C.I. for mean

Thus, the interval

$$\left(\bar{x} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

is called a $(1 - \alpha)$ 100 percent confidence interval for μ .

Remarks

- (i) The left endpoint is known as the lower confidence limit and the right endpoint is called the upper confidence limit.
- (ii) $(1 - \alpha)$ is called the confidence coefficient or the level of confidence.

Thus, for a normal population with known σ , a $(1 - \alpha)$ 100 percent confidence interval for mean is given as :

$$\bar{\mathbf{x}} - \mathbf{Z}_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{\mathbf{n}}} < \mu < \bar{\mathbf{x}} + \mathbf{Z}_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{\mathbf{n}}}$$

Example 7.5: Assume that the standard deviation of SAT verbal scores in a school system is known to be 100. A researcher wishes to estimate the mean SAT score and compute a 95% confidence interval from a random sample of 10 scores. The 10 scores are: 320, 380, 400, 420, 500, 520, 600, 660, 720, and 780.

Solution:

Let X: denote the SAT verbal scores.

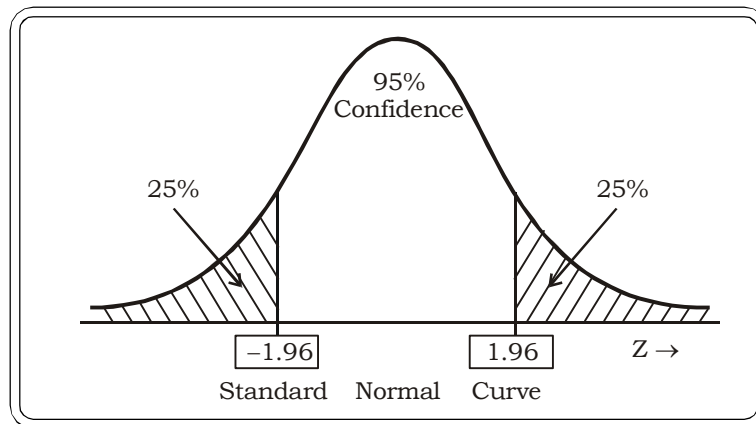
Here $\bar{x} = 530$, $n = 10$

Given: $\sigma = 100$

Thus $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 100/\sqrt{10} = 31.62$.

Let $\alpha = 0.05$

The value of z for the 95% confidence interval is the number of standard deviations one must go from the mean (in both directions) to contain 95% of the scores. It turns out that one must go 1.96 standard deviations from the mean in both directions to contain 0.95 of the scores. The value of 1.96 was found using a standard normal or z table. Since each tail is to contain 0.025 of the scores, we find the value of z for which $1 - 0.025 = 0.975$ of the scores are below. This value is 1.96.



All the components of the confidence interval are now known: $\bar{x} = 530$, $\sigma_{\bar{x}} = 31.62$, $z = 1.96$. Thus

$$\text{Lower limit} = 530 - (1.96)(31.62) = 468.02$$

$$\text{Upper limit} = 530 + (1.96)(31.62) = 591.98$$

Therefore the confidence interval is $(468.02 \leq \mu \leq 591.98)$. This means that the experimenter can be 95% certain that the mean SAT in the school system is between 468 and 592. Notice that this is a rather large range of scores. Naturally, if a larger sample size had been used, the range of scores would have been smaller.

The computation of the 99% confidence interval is exactly the same except that 2.58 rather than 1.96 is used for z . The 99% confidence interval is: $448.54 \leq \mu \leq 611.46$. As it must be, the 99% confidence interval is even wider than the 95% confidence interval.

Example 7.6: The production manager of a factory producing optical lenses wants to estimate the mean thickness of the lenses produced. A random sample of 50 lenses revealed a mean thickness of 0.50 mm. It is known that the population standard deviation is 0.15 mm.

Calculate a 98% confidence interval for the true mean thickness of the optical lenses produced.

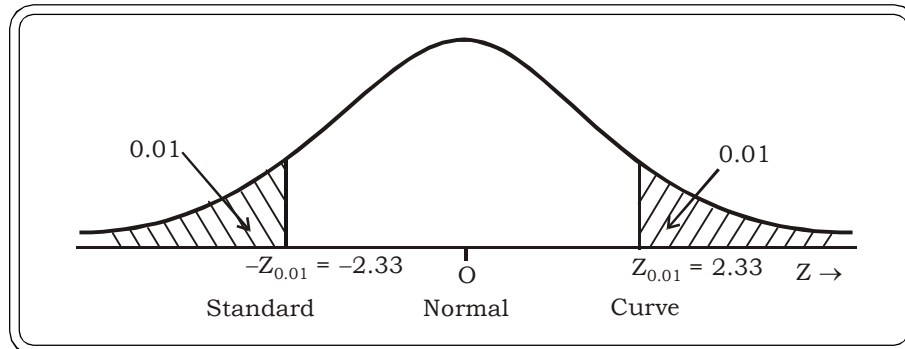
Solution:

Since we have to construct a 98% confidence interval,

$$1 - \alpha = 0.98$$

$$\Rightarrow \alpha = 0.2$$

$$\Rightarrow \frac{\alpha}{2} = 0.01$$



Thus, $Z_{0.01} = 2.33$ (From standard normal tables)

Population standard deviation: $\sigma = 0.15$ mm

Sample mean : $\bar{x} = 0.50$ m

Sample size : $n = 50$

Thus, a 98% confidence interval for the true mean thickness of the optical lenses produced is :

$$\begin{aligned} & \left(\bar{x} - Z_{0.01} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{0.01} \frac{\sigma}{\sqrt{n}} \right) \\ &= \left(0.50 - 2.33 \frac{0.15}{\sqrt{50}}, 0.50 + 2.33 \frac{0.15}{\sqrt{50}} \right) \\ &= (0.45, 0.55) \end{aligned}$$

Interpretation

The interval (0.45, 0.55) would include the true mean thickness of the optical lenses with probability 0.98.

Example 7.7: A sponsor of a television program targeted at the youth (age 16 years to 22 years) wants to find out the average amount of time young people spend watching television per week. A sample of 50 youth gave the average amount of time as 27.2 hours. From previous experience, the population standard deviation of the weekly extent of television watched is known to be 8 hours.

Construct a 99% confidence interval of the average amount of time young people spend watching television per week and state the conclusion.

Solution:

Let X : amount of time young people spend watching television per week

Sample information:

Sample size: $n = 50$

Sample mean: $\bar{x} = 27.2$ hours

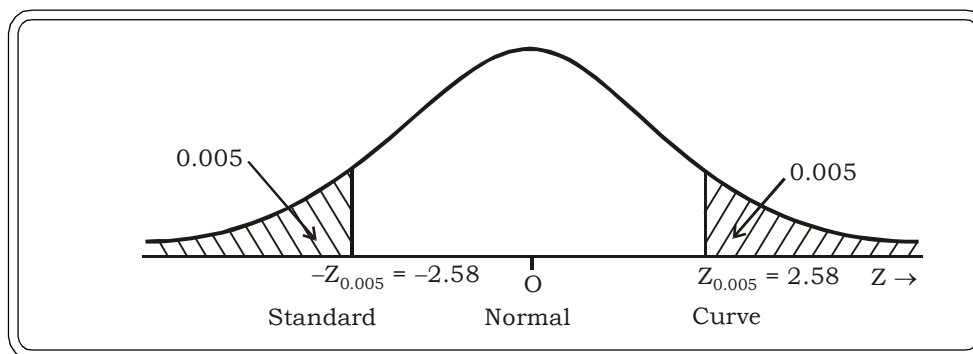
Population standard deviation is known to be $\sigma = 8$ hours

To construct a 99% confidence interval

$$1 - \alpha = 0.99$$

$$\Rightarrow \alpha = 0.01$$

$$\Rightarrow \frac{\alpha}{2} = 0.005$$



Thus, $Z_{0.005} = 2.58$ { From standard normal tables}

The 99% confidence interval for the average amount of time young people spend watching television is:

$$\begin{aligned} & \left(\bar{x} - Z_{0.005} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{0.005} \frac{\sigma}{\sqrt{n}} \right) \\ & = \left(27.2 - 2.58 \times \frac{8}{\sqrt{50}}, 27.2 + 2.58 \times \frac{8}{\sqrt{50}} \right) \\ & = (24.28, 31.12) \end{aligned}$$

Conclusion

With 99% level of confidence, it can be concluded that the average amount of time that a youth spends watching television is between 24.28 hours to 30.12 hours.

Confidence interval for μ when standard deviation is unknown and estimated

Assumptions:

1. The parent population follows a normal distribution.
2. Scores/observations are sampled randomly and are independent

It is very rare for a researcher wishing to estimate the mean of a population to already know its standard deviation. Therefore, the construction of a confidence interval almost always involves the estimation of both μ and σ .

(a) When σ is known, the formula

$(\bar{x} - z \times \sigma_{\bar{x}} < \mu < \bar{x} + z \times \sigma_{\bar{x}})$ is used for obtaining a confidence interval for the population mean.

(b) When σ , the population standard deviation is unknown, there are two cases:

- (i) Case I: When n , the sample size is large i.e. $n \geq 30$
- (ii) Case II: When $n < 30$ i.e. the sample size is small
 - (i) Case I: $n \geq 30$

In this case, σ -the population standard deviation (σ) can be approximated by the sample standard deviation (s).

Thus, the 100 (1 - α)% confidence interval for μ is

$$\left(\bar{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right)$$

$$\text{where } s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

(ii) Case II: $n < 30$

If σ is known, then the interval is:

$$\left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

If σ is unknown, a t-distribution is used instead of the z-distribution or the standard normal distribution.

The 100 (1 - α)% confidence interval for μ is:

$$\left(\bar{x} - t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}} \right)$$

where $t_{\frac{\alpha}{2}, n-1}$ is the table value for a t-distribution with $(n - 1)$ degrees of freedom such that area to the right of $t_{\frac{\alpha}{2}, n-1}$ is $\frac{\alpha}{2}$.

Also, since the t-distribution is symmetric, area to the left of $-t_{\frac{\alpha}{2}, n-1}$ is also $\frac{\alpha}{2}$.

Example 7.8: A sample of 64 students was able to finish a geometry test in an average time of 27.75 minutes with a standard deviation of 5.083 minutes.

Construct a 95% confidence intervals (C.I.) for the population mean.

Solution:

Since the sample size is large, the population standard deviation can be approximated by the sample standard deviation i.e.

$$\sigma = s = 5.083$$

Also, since a 95% C.I. is to be constructed, as shown before.

$$Z_{0.025} = 1.96$$

$$n = 64$$

$$\begin{aligned} \bar{X} &= \text{Average time taken by 64 students to complete the geometry test} \\ &= 27.75 \text{ minutes} \end{aligned}$$

The required C.I:

$$\begin{aligned} &= \left(\bar{x} - Z_{0.025} \frac{s}{\sqrt{n}}, \bar{x} + Z_{0.025} \frac{s}{\sqrt{n}} \right) \\ &= \left(27.75 - 1.96 \times \frac{5.083}{8}, 27.75 + 1.96 \times \frac{5.083}{8} \right) \\ &= (26.50, 28.99) \end{aligned}$$

Conclusion

With 95% level of confidence, it may be concluded that the average time students need to finish the geometry test is between 26.50 to 28.99 minutes.

Example 7.9: The manager of an insurance claims department wants to find out the average amount of money paid to claimants of automobile accidents. A study of 2000 claims paid out over a period of a year indicated that the average amount of money paid per claim was Rs. 1000 with a standard deviation of Rs. 200.

Construct a 90% confidence interval for the mean claim payment.

Solution:

Since, the sample size is large, we can use $s = \sigma$ and the normal approximation.

Let X: amount of money paid to a claimant.

Sample size $n = 2000$

Sample average amount: $\bar{x} = \text{Rs. } 1000$

Sample standard deviation: $s = \text{Rs. } 200$

Since we have to construct a 90% C.I.

$$1 - \alpha = 0.90$$

$$\Rightarrow \alpha = 0.10$$

$$\Rightarrow \frac{\alpha}{2} = 0.05$$

Thus $Z_{0.05} = 1.64$

The 90% C.I. for the mean is:

$$\begin{aligned} &= \left(\bar{x} - Z_{0.05} \frac{200}{\sqrt{2000}}, \bar{x} - Z_{0.05} \frac{200}{\sqrt{2000}} \right) \\ &= \left(1000 - 1.64 \times \frac{200}{44.72}, 1000 - 1.64 \frac{200}{44.72} \right) \\ &= (992.67, 1007.33) \end{aligned}$$

Conclusion

With 90% confidence, it may be concluded that the average amount of money paid to claimants of automobile accidents is between Rs. 992.67 and Rs. 1007.33.

Example 7.10: A machine filled 6 randomly picked paint cans with the following amounts of paints (in kgs).

15.7 15.9 16.2 16.3 15.8 15.9.

Set a 99% confidence interval for the true mean weight of the paint cans.

Solution:

Confidence level = 0.99

$$\Rightarrow 1 - \alpha = 0.99$$

$$\Rightarrow \frac{\alpha}{2} = 0.005$$

Since the sample size is small and the population standard deviation is unknown, we have to calculate the confidence interval based on t-distribution i.e.

$$\left(\bar{x} - t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}} \right)$$

We calculate \bar{x} and s ,

Sample size: $n = 6$

$$\begin{aligned}\text{Sample mean: } \bar{x} &= \frac{15.7 + 15.9 + 16.2 + 16.3 + 15.8 + 15.9}{6} \\ &= \frac{95.8}{6} \\ &= 15.97\end{aligned}$$

Sample standard deviation:

$$\begin{aligned}s^2 &= \frac{1}{6-1} [(15.7 - 15.97)^2 + (15.9 - 15.97)^2 + (16.2 - 15.97)^2 + (16.3 - 15.97)^2 \\ &\quad + (15.8 - 15.97)^2 + (15.9 - 15.97)^2] \\ &= \frac{1}{5} [0.0729 + 0.0049 + 0.0529 + 0.1089 + 0.0289 + 0.0049] \\ &= \frac{1}{5} [0.2734]\end{aligned}$$

$$s^2 = 0.05468$$

Thus $s = 0.2338$

$$t_{0.005,5} = 4.032$$

\therefore The required C.I. :

$$\begin{aligned}&\left(15.97 - 4.032 \times \frac{0.2338}{2.45}, 15.97 + 4.032 \times \frac{0.2338}{2.45} \right) \\ &= (15.58, 16.35)\end{aligned}$$

Conclusion

With 99% confidence level, it may be concluded that the true mean weight of the paint cans is between 15.58 kgs and 16.06 kgs.

Example: 7.11: Assume a researcher were interested in estimating the mean reading speed (number of words per minute) of high-school graduates and computing the 95% confidence interval. A sample of 6 graduates was taken and the reading speeds were: 200, 240, 300, 410, 450, and 600. Calculate the C.I.

Solution:

The sample mean: $\bar{x} = 366.6667$

The sample s.d: $s_{\bar{x}} = 60.9736$

Degrees of freedom: $df = 6 - 1 = 5$

$$t_{0.025,5} = 2.571$$

Therefore, the lower limit is: $\bar{x} - t \times s_{\bar{x}} = 296.69$ and the upper limit is: $\bar{x} + t \times s_{\bar{x}} = 436.65$.
Therefore, the 95% confidence interval is:

$$296.69 \leq \mu \leq 436.65$$

Thus, the researcher can be 95% sure that the mean reading speed of high-school graduates is between 296.69 and 436.65 words per minute.

Example 7.12: A particular branch of a large multinational bank is trying to estimate the mean amount of time (in minutes) that a customer care executive spends on a customer. The bank manager samples 5 customers and records the amount of time the executive spends on each customer. The sample mean time came out to be 3.2 minutes with a sample standard deviation of 1.2 minutes.

Estimate a 90 % confidence interval for the mean amount of time an executive spends on a customer.

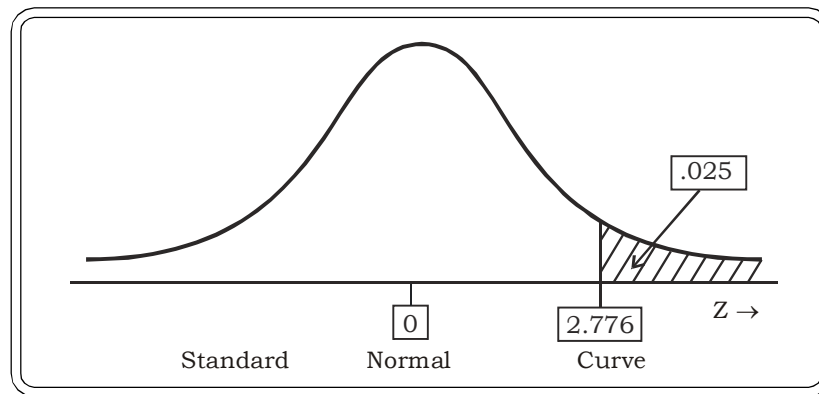
Solution:

$$n = 5 \text{ customers}$$

$$\bar{X} = 3.2 \text{ minutes}$$

$$s = 1.2 \text{ minutes}$$

$$t_{0.025,4} = 2.776$$



The 90 % confidence interval for the mean amount of time is

$$\begin{aligned} [3.2 \pm (2.776) 1.2/\sqrt{5}] &= 3.2 \pm 2.776(0.54) \\ &= \{3.2 - 1.49, 3.2 + 1.49\} = \{1.71, 4.69\} \end{aligned}$$

Conclusion

With 90% confidence, we can conclude that the mean amount of time an executive spends on a customer is between 1.71 minutes to 4.69 minutes.

Example 7.13: A manufacturer of cells tested 25 cells to find their mean lifetime. The sample indicated an average of 420 hours with a standard deviation of 43 hours. Find a 95% confidence interval for mean lifetime of the cells.

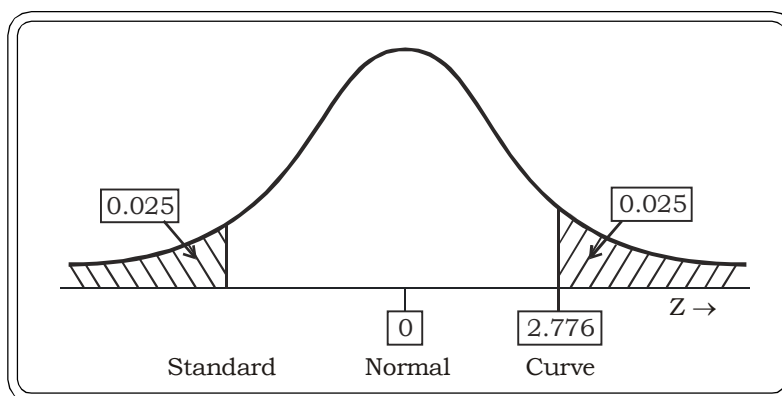
Solution:

$$n = 25$$

$$\bar{X} = 420 \text{ hours}$$

$$s = 43 \text{ hours}$$

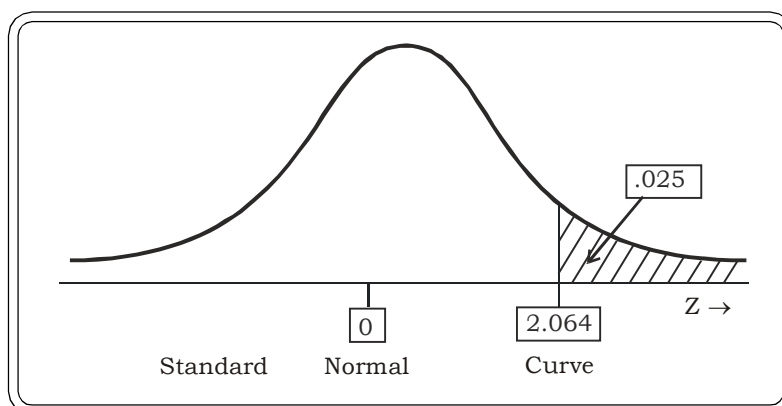
$$t_{\frac{\alpha}{2},24} = t_{0.025,24} = 2.064$$



95% confidence interval of mean lifetime of the cells

$$\left[\bar{x} - t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}} \right]$$

$$= 420 \pm t_{24, 0.025} \frac{43}{\sqrt{25}}$$



$$= [420 - 2.064 (8.6), 420 + 2.064 (8.6)]$$

$$= [420 - 17.75, 420 + 17.75]$$

$$= [402.25, 437.75]$$

With 95% confidence, we can say that the mean lifetime of the cells is between 402.25 hours to 437.75 hours.

7.2.2.2 Interval Estimator of Difference of Two Means

The estimators discussed so far are concerning a single population. However, on many occasions it is desired to compare parameters of two populations or proportions of a given attribute in two populations. These parameters could be means of two populations or proportions of a given attribute in two populations. In this case, we take random samples from each of the two populations to be compared and then compare the estimates obtained from the sample. In this section, we discuss comparison of two population means.

Consider two populations, each of which is assumed to be normally distributed with means μ_1 and μ_2 . Our objective is to set a confidence interval for $\mu_1 - \mu_2$.

A general format for confidence interval is

Point Estimator \pm Maximum Error

= Point Estimator \pm Distribution Value \times Standard Error

The Point Estimator for $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2$ and may be computed as follows:

With known variances

If \bar{x}_1 and \bar{x}_2 are means of two independent random samples of size n_1 and n_2 from approximately normal populations **with known variances**, σ_1^2 and σ_2^2 respectively, a $(1 - \alpha)$ 100% confidence interval for $\mu_1 - \mu_2$ is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $z_{\frac{\alpha}{2}}$ is the z value leaving an area of $\alpha/2$ to the right.

If n_1 and n_2 are large, and s_1^2 and s_2^2 represent sample estimates of the variances, then $(1 - \alpha)$ 100% C.I. is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

With unknown but equal variances

If \bar{x}_1 and \bar{x}_2 are means of two independent random samples of size n_1 and n_2 from approximately normal populations **with unknown but equal variances**, an approximate $(1 - \alpha)$ 100% confidence interval for $\mu_1 - \mu_2$ is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, s_p} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $t_{\frac{\alpha}{2}, v}$ is the t value with $v = n_1 + n_2 - 2$ degrees of freedom, leaving an area of $\alpha/2$ to the right. s_p is the pooled estimate of the population standard deviation that can be calculated as follows:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

With unknown and unequal variances

If \bar{x}_1 and s_1^2 , and \bar{x}_2 and s_2^2 are the means and variances of independent random samples of size n_1 and n_2 from approximately normal populations with unknown and unequal variances, an approximate $(1 - \alpha)$ 100% confidence interval for $\mu_1 - \mu_2$ is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $t_{\alpha/2}$ is the t value with v degrees of freedom, leaving an area of $\alpha/2$ to the right. Here v is defined as

$$v = n_1 + n_2 - 2$$

Example 7.14: Consider two normally distributed populations. The first population has variance 10. A sample of size 30 was selected from this population and its mean was found out to be 8. The second population variance is 12 and a sample of size 35 gave a mean of 7. Assuming that the samples were drawn independently, estimate a 95% confidence interval for $\mu_1 - \mu_2$.

Solution:

Ist Population:

$$\sigma_1^2 = 10$$

$$n_1 = 30$$

$$\bar{x}_1 = 8$$

IInd Population

$$\sigma_2^2 = 12$$

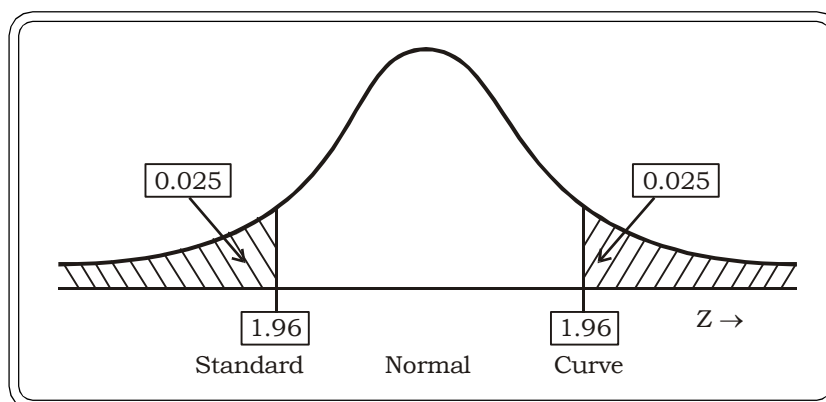
$$n_2 = 35$$

$$\bar{x}_2 = 7$$

95% confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$Z_{\frac{0.05}{2}} = 1.96$$



$$= (1 - 1.61, 1 + 1.61)$$

$$= (-0.61, 2.61)$$

Example 7.15: In 10 half an hour morning programs, the mean time devoted to commercials was 6.8 minutes with $s_1^2 = 1$ minute. In 12 half hour evening programs, the mean time was 5.6 minutes with $s_1^2 = 1.3$ minutes. Estimate the difference in the true mean times devoted to commercials during the morning and evening half an hour programs, using a 90 percent confidence interval. The variance of the time devoted to commercials in the morning and evening programs are assumed equal.

Solution:

For morning programs:

$$\bar{x}_1 = 6.8$$

$$n_1 = 10$$

$$s_1^2 = 1 \text{ minute}$$

For evening programs:

$$\bar{x}_2 = 5.6$$

$$n_2 = 12$$

$$s_2^2 = 1.3 \text{ minute}$$

The 90% confidence interval for the difference in the true mean times devoted to commercials is

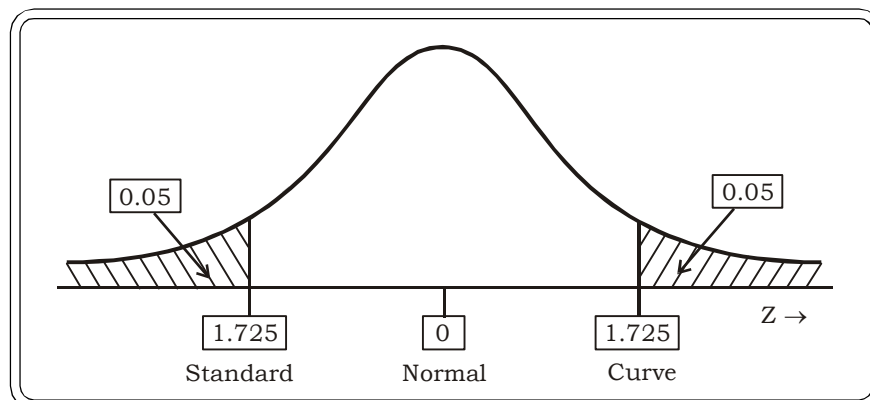
$$(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}, n_1 + n_2 - 2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

$$s^2 = \frac{9 \times 1 + 11 \times 1.3}{20} = 1.165$$

$$\alpha = 0.10$$

$$t_{.05, 20} = 1.725$$



Thus the C.I is:

$$\begin{aligned} & [1.2 \pm (1.725) (1.165) \sqrt{0.1+0.08}] \\ & = [1.2 \pm (2.01) (0.42)] \\ & = [1.2 \pm 0.84] \\ & = [0.36, 2.04] \end{aligned}$$

7.2.2.3 Interval Estimator of Single Population Proportion

Assumptions

1. The sample is a simple random sample.
2. The conditions for the binomial distribution apply: These are
 - (i) There are a fixed number of trials
 - (ii) The trials are independent
 - (iii) There are two categories of outcomes, success and failure
 - (iv) The probability of success remain constant for each trial.

The probability or proportion of success and failure in the population are denoted by P and Q respectively. Since P and Q are not known, we use the sample proportion to estimate their values.

The estimated sample proportion \hat{p} is

$$\hat{p} = \frac{x}{n} = \text{Sample proportion (of } x \text{ successes in a sample of size } n) \text{ i.e. the point estimate of } p.$$

Thus the estimated q i.e. $\hat{q} = 1 - \hat{p}$

Thus, an approximate $(1 - \alpha)100$ percent confidence interval for the population proportion P is given by

$$\frac{x}{n} - z_{\frac{\alpha}{2}} \sqrt{\frac{\frac{x}{n} \left(1 - \frac{x}{n}\right)}{n}} < p < \frac{x}{n} + z_{\frac{\alpha}{2}} \sqrt{\frac{\frac{x}{n} \left(1 - \frac{x}{n}\right)}{n}}$$

provided n is large i.e. $n \geq 30$, approximately.

The interval may also be written as

$$\hat{p} - E < P < \hat{p} + E$$

$$\text{where } \hat{p} = \frac{x}{n}, \hat{q} = 1 - \frac{x}{n} \text{ \& } E = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} = \text{Margin of Error}$$

Example 7.16: In an attempt to control the quality of output of cathode tubes in a factory, a sample of parts is chosen randomly and examined to estimate the proportion of defective tubes. In a random sample of 100 tubes, 12 were found to be defective.

- (i) Determine the margin of error at 98% confidence level.
 (ii) Calculate a 98% confidence interval of population proportion defective.

Solution:

Sample size: $n = 100$

$$\begin{aligned} X &\rightarrow \text{no of defective tubes in the sample} \\ &= 12 \end{aligned}$$

Sample proportion of defectives : $\hat{p} = \frac{12}{100} = 0.12$

- (i) For 98% confidence level

$$1 - \alpha = 0.98$$

$$\Rightarrow \frac{\alpha}{2} = 0.01$$

and $Z_{0.01} = 2.33$

$$\begin{aligned} \text{Thus, margin of error} &= Z_{0.01} \sqrt{\frac{\hat{p}\hat{q}}{n}} \\ &= 2.33 \sqrt{\frac{(0.12)(0.88)}{100}} \\ &= 2.33 (0.032) \\ &= 0.076 \end{aligned}$$

- (ii) A 98% C.I. of population proportion defective is:

$$\begin{aligned} &= (\hat{p} - E, \hat{p} + E) \\ &= (0.12 - 0.076, 0.12 + 0.076) \\ &= (0.044, 0.196) \end{aligned}$$

Thus, proportion of defective tubes produced is between 4% to 19%.

Example 7.17: In a sample of 750 people, 27% said they feel that health care is the most important issue facing our country. Estimate an interval for the proportion of people who feel that health care is the most important issue facing our country.

Solution:

Here the sample proportion is

$$\hat{p} = 0.27$$

Assuming 95% confidence level, so $z_{\alpha/2} = 1.96$ (from standard normal tables)

As we know, the interval estimator of single proportion is given by

$$\hat{p} - E < p < \hat{p} + E$$

Now

$$E = Z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96 \sqrt{\frac{(0.27)(0.73)}{750}} = .0318$$

$$\text{Lower Limit : } \hat{p} - E = 0.27 - .0318 = 0.2382$$

$$\text{Upper Limit : } \hat{p} + E = 0.27 + .0318 = 0.3018$$

So our confidence interval is $0.238 < p < 0.302$.

Conclusion

This implies that between 23.8% to 30.2 % people feel that health care is the most important issue for the country.

Example 7.18: A cosmetic company launched a new brand of nail polish by advertising on different T.V. channels. To find out the percentage of people who had seen the advertisement a random sample of 100 people were questioned. Out of these 60 responded in the affirmative. Estimate a 90% confidence interval for the true proportion of people who had seen the advertisement.

Solution:

$$n = 100$$

$$x = 60 = \text{Number of people who had seen the advertisement}$$

$$p = 0.60 = \text{Proportion of people in the sample who had seen the advertisement.}$$

$$\sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.60)(0.40)}{100}} = 0.05$$

$$\begin{aligned} \text{The 90\% C.I.} &= \left[p \pm Z_{\frac{.05}{2}} \sqrt{\frac{pq}{n}} \right] \\ &= [0.60 \pm (1.96)(0.05)] \\ &= [0.60 \pm 0.098] \\ &= [0.502, 0.698] \end{aligned}$$

Conclusion

With 90% confidence level, we may conclude that the proportion of people who had seen the advertisement is between 50% to 69.8%.

7.2.2.4 Interval Estimator of Difference of Two Population Proportions

Suppose we wish to compare the proportion of women favoring public transport to the proportion of men favoring public transport. One population would consist of the collection of all men and the other the collection of all women. If P_1 represents the proportion of women in favor of public transport and P_2 represents the proportion of men in favor of public transport, then our intention is to construct a confidence interval for $P_1 - P_2$. Let n_1 and n_2 be the sample sizes from the two populations and x_1 be the number of women in the sample in favor of public transport and x_2 the number of men in the sample in favor of public transport. Then

$(\hat{p}_1 - \hat{p}_2) = \frac{x_1}{n_1} - \frac{x_2}{n_2}$ will be a point estimator of $P_1 - P_2$.

$$\hat{q}_1 = 1 - \frac{x_1}{n_1} \quad \text{and} \quad \hat{q}_2 = 1 - \frac{x_2}{n_2}$$

For sufficiently large sample sizes, n_1 and n_2 , the sample distribution of $(\hat{p}_1 - \hat{p}_2)$, based on independent random samples from two populations, is approximately normal with

Mean: $\mu_{(\hat{p}_1 - \hat{p}_2)} = (\hat{p}_1 - \hat{p}_2)$

and

Standard deviation: $\sigma_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$

A $(1 - \alpha)$ 100 percent confidence interval for $p_1 - p_2$ would be:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\frac{\alpha}{2}} \sigma_{(\hat{p}_1 - \hat{p}_2)} \cong (\hat{p}_1 - \hat{p}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} = (\hat{p}_1 - \hat{p}_2) \pm E$$

where $E = \text{margin of error} = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$

Assumption 1: The samples are sufficiently large so that the approximation is valid. As a general rule of thumb we will require that intervals

$$\hat{p}_1 \pm 2 \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1}} \quad \text{and} \quad \hat{p}_2 \pm 2 \sqrt{\frac{\hat{p}_2 \hat{q}_2}{n_2}} \quad \text{do not contain 0 or 1.}$$

Assumption 2: Neither p_1 nor p_2 is close to zero or 1.

Example 7.19: Suppose that there were two surveys, one was carried out in 2000 and another in 2006. In both surveys, random samples of 1,400 adults in a country were asked whether they were satisfied with their life. The results of the surveys are reported in the table below. Construct a point estimate for difference between the proportion of adults in the country in 2000 and in 2006 who were satisfied with their life.

Proportions of two samples:

Year	2000	2006
Number surveyed	$n_1 = 1,400$	$n_2 = 1,400$
Number in sample who said they were satisfied with their life	462	674

Estimate a confidence interval for the difference between the proportions of the adults in this country in 2000 and in 2006 who said that they were satisfied with their life, using a 95% confidence interval.

Solution:

p_1 = Population proportion of adults who said that they were satisfied with their life in 2000.

p_2 = Population proportion of adults who said that they were satisfied with their life in 2006.

To judge the reliability of the point estimate $(\hat{p}_1 - \hat{p}_2)$, we need to know the characteristics of its performance in repeated independent sampling from two populations. This information is provided by the sampling distribution of $(\hat{p}_1 - \hat{p}_2)$.

As a point estimate of $(p_1 - p_2)$, we will use the difference between the corresponding sample proportions, $(\hat{p}_1 - \hat{p}_2)$, where

$$\hat{p}_1 = \text{Proportion of satisfied adults in 2000} = \frac{462}{1400} = 0.33$$

$$\text{and } \hat{p}_2 = \text{Proportion of satisfied adults in 2006} = \frac{674}{1400} = 0.48$$

Thus, the point estimate of $(p_1 - p_2)$, is

$$(\hat{p}_1 - \hat{p}_2) = 0.33 - 0.48 = -0.15$$

where $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

Thus $q_1 = 0.67$ and $q_2 = 0.52$.

$$\begin{aligned} E &= \text{margin of error} = z_{0.025} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \\ &= 1.96 \sqrt{\frac{(0.33)(0.67)}{1400} + \frac{(0.48)(0.52)}{1400}} \\ &= 1.96 \sqrt{0.00016 + 0.00018} \\ &= 0.036 \end{aligned}$$

Thus, the 95% C.I. for the difference in two population proportions is:

$$\begin{aligned} &(-0.15 - 0.036, -0.15 + 0.036) \\ &= (-0.186, -0.114) \end{aligned}$$

Example 7.20: In a survey conducted in a metro, 50 out of 150 student respondents liked a new advertisement and 160 out of 300 working respondents liked the new advertisement.

- (i) Obtain a point estimate of the difference in proportion of students & working respondents who liked the new advertisement.
- (ii) Construct a 95% C.I. for $P_1 - P_2$.

Solution:

Given: $x_1 = 50$ (Number of students who liked the advertisement)

$x_2 = 160$ (Number of working respondents who liked the advertisement)

$n_1 = 150$ (Sample size of students)

$n_2 = 300$ (Sample size of working respondents)

$$\text{Here } \hat{p}_1 = \frac{50}{150}; \hat{p}_2 = \frac{160}{300}; \hat{q}_1 = \frac{100}{150}; \hat{q}_2 = \frac{140}{300}$$

Thus,

$$\hat{p}_1 = 0.33; \hat{p}_2 = 0.53 \quad \hat{q}_1 = 0.67; \hat{q}_2 = 0.47$$

A point estimate of the difference in the proportion of students and working respondents who liked the new advertisement is:

$$\hat{p}_1 - \hat{p}_2 = -0.2$$

A 95% confidence Interval for difference of population proportions is given by:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} = \hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}} \sigma_{\hat{p}_1 - \hat{p}_2}$$

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{(0.33)(0.67)}{150} + \frac{(0.53)(0.47)}{300}}$$

$$\sigma_{\hat{p}_1 - \hat{p}_2} = 0.05$$

$$\& \quad z_{\frac{0.05}{2}} = 1.96 \quad (\text{from standard normal tables})$$

Thus the 95% confidence interval for difference in population proportions is calculated as

$$= -0.2 \pm 1.96(0.050)$$

$$= -0.2 \pm 0.098$$

$$= (-0.2 - 0.098, -0.2 + 0.098)$$

$$= (-0.298, -0.102)$$

7.2.2.5 Determination of Sample Size

In this section we discuss determination of sample size for two cases viz.

(i) Case I: Sample size determination for estimating population mean

Suppose we wish to determine how large a sample we must take in order to be $(1 - \alpha)100$ percent confident that the sample mean \bar{x} would not differ from the population mean μ by some given amount.

In practice, let

$$E = \bar{x} - \mu$$

Then, E is called the sampling error or margin of error or error in estimation.

As an example, suppose a quality engineer knows that the breaking strength of cables have a normal distribution with a standard deviation $\sigma = 5$ pounds. He wants to be 90 percent confident that the sample mean of a random sample he considers should not differ from the true mean breaking strength by more than 0.75 pounds.

$$\text{Thus, } \bar{x} - \mu \leq 0.75$$

0.75 is called the maximum error and is given by:

$$e = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

On simplifying this expression:

$$\Rightarrow n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{e} \right)^2$$

Thus, the sample size, so as to be $(1 - \alpha)100$ percent confident that the estimate \bar{x} does not differ from the true mean μ by a quantity e , which is preassigned is:

$$n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{e} \right)^2$$

Example 7.21: Suppose the mean idle time of a machine is to be estimated within 1.15 hour of the true mean idle time with 98% level of confidence. It is known from past data, that the idle time of a machine is normally distributed with a standard deviation of 2 hours. Compute the appropriate sample size.

Solution:

Margin of error = 1.15 hrs.

Level of confidence = 0.98

$$\Rightarrow \alpha = 0.02$$

$$\Rightarrow z_{\frac{\alpha}{2}} = z_{0.01} = 2.33$$

Standard deviation: $\sigma = 2$ hours

Thus, the sample size:

$$\begin{aligned} n &= \left(\frac{z_{\frac{\alpha}{2}} \sigma}{e} \right)^2 \\ &= \left(\frac{2.33 \times 2}{1.15} \right)^2 \\ &\cong 16 \end{aligned}$$

Thus, an appropriate sample size = 16

Case II : Sample size determination for estimating population proportion

In estimating a proportion also, it may be desirable to know the right sample size to consider. For example: a researcher may want to know how many households he needs to interview to be 90% confident that his estimate of proportion of people who owns a LCD television will not differ from the true proportion P by say, not more than 0.01.

An approximate estimate of the sample size so as to be $(1 - \alpha)$ 100 percent confident that the estimate of a sample proportion $(p = \frac{x}{n})$ would not differ from the true population proportion by more than a quantity e is:

$$n = \frac{z_{\frac{\alpha}{2}}^2}{4e^2}$$

Example 7.22: A manufacturer of watches wants to estimate the proportion of defective watches produced in the factory. He wants to be 95% confident that his estimate would not differ from the true proportion of defectives by more than 0.02. How large a sample should he consider?

Solution:

Confidence level: $1 - \alpha = 0.95$

$$\Rightarrow \alpha = 0.05$$

$$\Rightarrow \frac{\alpha}{2} = 0.025$$

$$z_{0.025} = 1.96$$

Margin of error acceptable : $e = 0.02$

Thus, the sample size may be estimated by the formula:

$$\begin{aligned} n &= \frac{z_{0.025}^2}{4e^2} = \frac{(1.96)^2}{4(0.02)^2} \\ &= \frac{3.8416}{0.0016} = 2401 \end{aligned}$$

7.3 TESTING OF HYPOTHESIS

In the above section, we considered one aspect of statistical inference. Testing of hypothesis is another essential part of statistical inference. In order to formulate a test, usually, some theory has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. A Hypothesis is nothing but the assumption to be tested. All hypothesis testing begins with a hypothesis. In seeking to learn more about the social world, social scientists ask many different kinds of questions about relationships between factors of social life. How do investors change their behavior when market conditions change? To address these questions, it is necessary to form hypotheses which can then be evaluated using sample data and finally lead to a decision of accepting or rejecting the hypothesis.

7.3.1 Null and Alternative Hypothesis

Generally speaking, two competing hypotheses are evaluated in the light of some empirical data. These hypotheses are referred to as the null hypothesis and the alternative hypothesis. The primary purpose of hypothesis testing is to examine the likelihood of the null hypothesis holding true with data. Now the question arises as to what is null and alternative hypothesis?

Null Hypothesis: Null hypothesis relates to the statement being tested. It represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument. Usually it is denoted by the symbol H_0 .

Alternative hypothesis: Any hypothesis, which is complementary to null hypothesis, is called alternative hypothesis. Alternative hypothesis is normally denoted by H_1 .

For example, while testing if a coin is fair, the null hypothesis would be

$H_0 : p = \frac{1}{2}$ i.e. the coin is fair and

the alternative hypothesis $H_1 : p \neq 1/2$: i.e. the coin is not fair.

Depending on the data, the null hypothesis either would or would not be rejected as a viable possibility. Specific criteria used to accept or reject the null hypothesis are discussed in the subsequent sections of this chapter.

Now, it is clear that the goal of any hypothesis testing is to make a decision. In particular, we will decide whether to reject the null hypothesis H_0 , in favor of the alternative hypothesis H_1 . Although we would always like to be able to make a correct decision, we must remember that the decision will be based on sample information, and thus we are subject to make one or two types of error, as defined in the following.

- (i) Decision 1: Accept H_0 when it is actually false
- (ii) Decision 2: Reject H_0 as false when it is actually true

The two other possible decisions which would lead us to correct decisions are:

- (i) Decision 3: Accept H_0 as true, when it is actually true.
- (ii) Decision 4: Reject H_0 as false, when it is actually false.

Critical Region and Level of Significance

The critical region, or rejection region, is a set of values of the test statistic for which the null hypothesis is rejected in a hypothesis test; that is, the sample space for the test statistic is partitioned into two regions; one region (the critical region) will lead us to reject the null hypothesis H_0 , the other to accept H_0 . So, if the observed value of the test statistic is a member of the critical region, we conclude 'reject H_0 '; if it is not a member of the critical region then we conclude 'do not reject H_0 '.

The critical region is called the rejection region and the other region is called the acceptance region.

The Test Statistic

The decision to accept or reject the null hypothesis is based on a rule or a procedure based on a statistic which is called the test statistic.

Level of Significance

The purpose of hypothesis testing is not to question the computed value of the sample statistic but to make a judgment about the difference between the sample statistic and hypothesized population parameter. The significance level is usually denoted by α .

The critical value, based on the level of significance is the value that separates the acceptance region and the rejection region. It determines how much difference between the sample statistic and the hypothesized population parameter may be considered as significant so as to reject the null hypothesis of no difference.

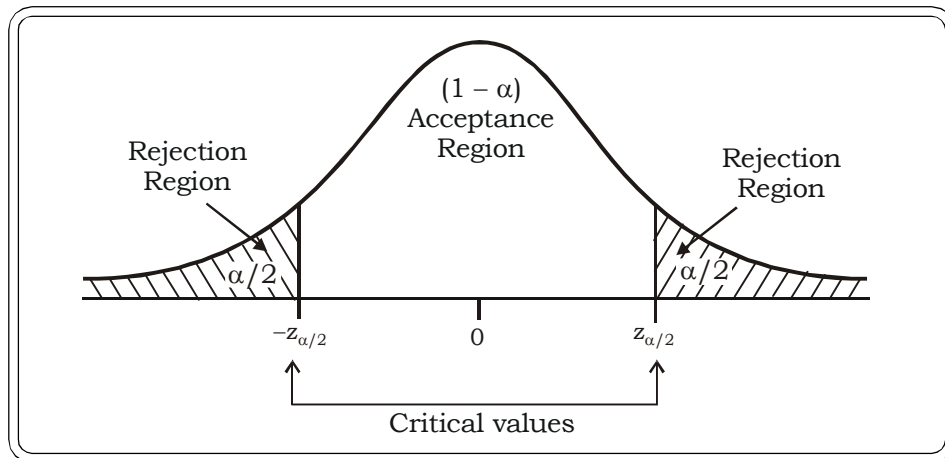


Figure 7.6

Critical Region and Critical Values

In the figure above, if the test statistic falls in the acceptance region, we may accept H_0 , and if it falls in the rejection region, we may reject H_0 .

7.3.2 Type I and Type II Error

In a hypothesis testing problem, there are two possibilities regarding the null hypothesis

- (a) H_0 is true
- (b) H_0 is false

(a) H_0 is true:

Type I Error

When H_0 is true, we can make two decisions

- (i) Accept H_0 or
- (ii) Reject H_0

The first decision leads us to a correct conclusion. The second decision would lead to a incorrect conclusion.

In a hypothesis test, type I error denoted by α , occurs when the null hypothesis is rejected when it is in fact true; that is, H_0 is wrongly rejected. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on an average than the current drug; that is H_0 : there is no significant difference between the two drugs on an average.

A type I error would occur if we concluded that the two drugs produced different effects when in fact there was no difference between them.

(b) H_0 is false:**Type II Error**

If H_0 is false, again, two possible decisions are

- (i) Accept H_0
- (ii) Reject H_0

The second decision leads to a correct conclusion the first decision leads to a wrong conclusion.

In hypothesis testing, type II error, denoted by β occurs, when the null hypothesis H_0 is accepted when it is in fact false. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on an average, than the current drug; that is H_0 : there is no difference between the two drugs on an average.

A type II error would occur if it was concluded that the two drugs produced the same effect, that is, there is no difference between the two drugs on average, when in fact they produced different ones. Type II error is frequently due to sample sizes being too small.

Producer's and Consumer's risk

Type I error is also known as **Producer's risk**, which indicates rejecting a good lot. For example, a manufacturer of pens rejects a lot of high quality pens due to standards that fall outside of their allowable range.

Type II error is also termed as a **Consumer's risk** and indicates accepting a bad lot. For example, a house is purchased that is believed to be of high quality but within a month the plumbing has failed. This is the risk a consumer has to face. The following table gives a summary of possible results of any hypothesis test:

Table 7.2
Type I and Type II Errors

		Decision	
		Reject H_0	Don't reject H_0
"State of Nature"	H_0	Type I Error (α) (Producer's Risk)	Right Decision ($1 - \alpha$)
	H_1	Right Decision ($1 - \beta$)	Type II Error (β) (Consumer's Risk)

A type I error is often considered to be more serious, and therefore more important to avoid, than a type II error. The hypothesis test procedure is therefore adjusted so that there is a guaranteed 'low' probability of rejecting the null hypothesis wrongly; this probability is never 0.

7.3.3 One-Tailed Test - Two Tailed Test

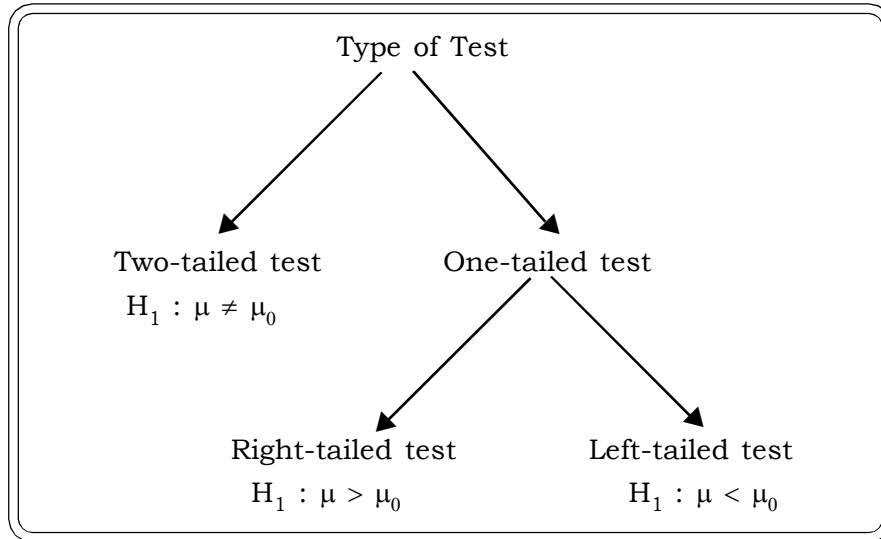


Figure 7.7

Types of Test

Two-Tailed Test

A two-sided test is a statistical hypothesis test in which the values for which we can reject the null hypothesis H_0 are located in both tails of the probability distribution. In other words, the critical region for a two-sided test is the set of values less than a first critical value of the test and the set of values greater than a second critical value of the test. A two-sided test is also referred to as a two-tailed test of significance.

For a two tailed test, the alternative hypothesis is usually of the form

$$H_1 : \mu \neq \mu_0$$

If level of significance is α %, then the two rejection regions equal to $\frac{\alpha}{2}$ % lie on each tail of the curve of the sampling distribution.

An Example: For a two tailed test, the null hypothesis and the alternative hypothesis is of the form.

$$H_0 : \mu = 50 \text{ i.e. the average number of matches in a box is } 50$$

$$H_1 : \mu \neq 50 \text{ i.e. the average number of matches in a box is not } 50$$

The alternative hypothesis says that the average is not equal to 50. It does not specify if the average is less than or greater than the mean.

Thus, nothing specific can be said about the average number of matches in a box and only that, if we reject the null hypothesis in our test, we would know that the average number of matches in a box is likely to be either less than or greater than 50. This is an example of a two tailed test.

One-Tailed Test

A one-sided test is a statistical hypothesis test in which the values for which we can reject the null hypothesis, H_0 are located entirely in one tail of the probability distribution. In other words, the critical region for a one-sided test is the set of values less than the critical value of the test, or the set of values greater than the critical value of the test. A one-sided test is also referred to as a one-tailed test of significance.

The choice between a one-sided and a two-sided test is determined by the purpose of the investigation or prior reasons for using a one-sided test.

For one tailed test, the alternative hypothesis may be of the form

(i) $H_1 : \mu > \mu_0$ or

(ii) $H_1 : \mu < \mu_0$

If level of significance is $\alpha\%$ then the rejection region is equal to $\alpha\%$ and lies on only one tail of the curve of the sampling distribution.

(i) In this case, the rejection region lies entirely on the right tail. The decision rule is to reject H_0 if the value of the statistic lies in the rejection region.

(ii) In this case, the rejection region lies entirely on the left tail.

An Example: Suppose we wanted to test a manufacturers claim that there are on an average 50 matches in a box. We could set up the following hypotheses:

$H_0: \mu = 50$ against

Either

$H_{01}: \mu > 50$ or $H_{02}: \mu < 50$

Either of these two alternative hypotheses would lead to a one-sided test.

H_{01} is a right tailed hypothesis and its rejection region will lie on the right side of the curve

H_{02} is a left tailed hypothesis and its rejection area will lie in the left side of the curve.

Presumably, from the consumer point of view, in this example we would want to test the null hypothesis against the first alternative hypothesis since it would be useful to know if there is likely to be less than 50 matches, on average, in a box (no one would complain if they get the correct number of matches in a box or more). And from the manufacturer's point of view it is critical that the average number of matches in a box should not exceed 50.

Identification of Rejection and Acceptance Regions under different tail

Figure 7.8 illustrates how to select and interpret the critical region under a two tail test. 5% level of significance indicates 95% area of curve is under the acceptance of null hypothesis. The two shaded areas representing a total of 5% area are the rejection regions. This implies if the test statistic falls in these two parts the null hypothesis will be rejected.

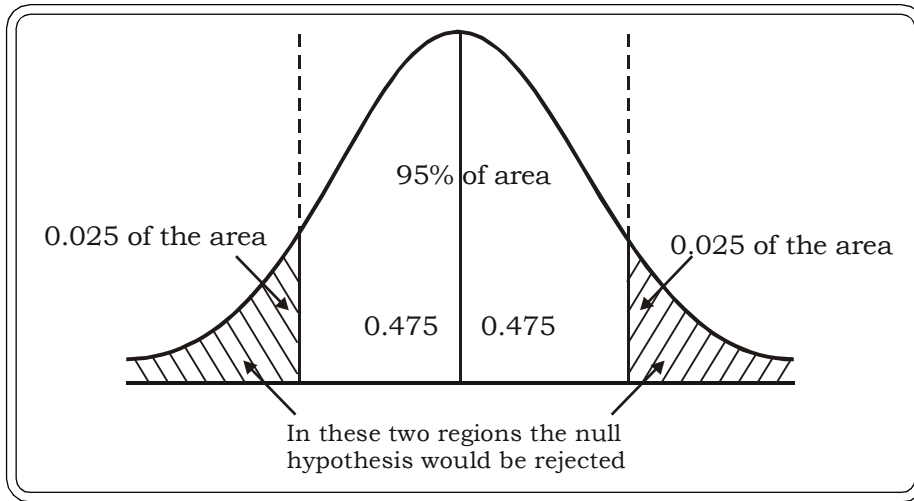


Figure 7.8

Two Tailed Test (5% level of significance)

Two-Tailed Test (5% level of significance)

In these two regions, the null hypothesis would be rejected

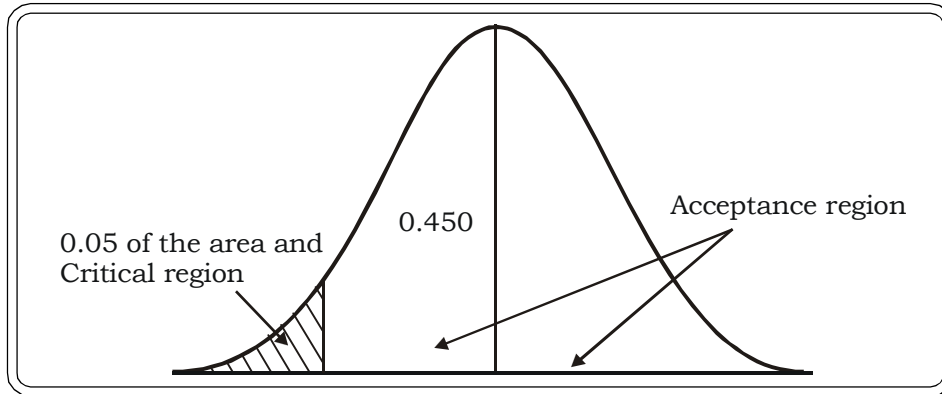


Figure 7.9

Left-Tailed Test (5% level of significance)

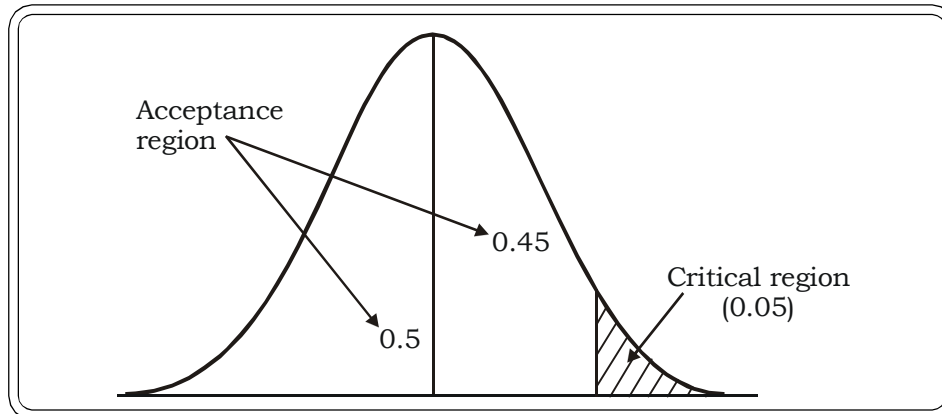


Figure 7.10

Right-Tailed Test (5% level of significance)

In this respect one important point to be noted here is that in each example of hypothesis testing when we accept the null hypothesis on the basis of sample information, we are saying that there is no statistical evidence to reject it. We are not saying that the null hypothesis is true. The only way to prove a null hypothesis is to know the population parameter, which is not possible with the sampling method.

Summary of Steps for Testing of Hypothesis

- Step 1. State the null and alternative hypothesis.
- Step 2. Establish a level of significance.
- Step 3. Select suitable test statistic.
- Step 4. Calculation of the test statistic.
- Step 5. Making decision by comparing with tabulated values.

If calculated value of statistic < Tabulated value of statistic we may accept the null hypothesis, else reject it at the assumed level of significance.

7.3.4 One Sample Tests

7.3.4.1 One Sample Z Test for Mean

The **Z-test** is a statistical test used in inference which determines if the difference between a sample mean and the population mean is sufficiently different as to be statistically significant.

For the Z-Test to be reliable, certain conditions must be met. The most important is that since the Z-Test uses the population mean and population standard deviation, so these must be known. The sample selected must be a simple random sample of the population. If the sample came from a different sampling method, a different formula must be used. It must also be known that the population values are distributed normally (i.e., the sampling distribution of the probabilities of possible values fits a standard normal curve). If it is not known that the population varies normally, it suffices to have a sufficiently large sample, generally agreed to be ≥ 30 or 40.

In practice, knowing the true σ of a population is unrealistic except for cases such as standardized testing in which the entire population is known. In cases where it is impossible to measure every member of a population it is more realistic to use a t-test, provided the sample size is small, which uses the standard error obtained from the sample along with the t-distribution.

Hypothesis testing of population means when population standard deviation is known

Summarizing, the test requires the following quantities to be known:

- σ (the standard deviation of the population)
- μ (the mean of the population)
- \bar{x} (the mean of the sample)
- n (the size of the sample)

First, calculate the standard error (SE) of the mean:

$$SE = \frac{\sigma}{\sqrt{n}}$$

The test statistic or the formula for calculating the z score for the Z Test is as follows:

$$Z = \frac{\bar{x} - \mu}{SE}$$

and $Z \sim N(0, 1)$

Finally, the calculated z score is compared to a tabulated z score, a table which contains the percent of area under the normal curve between the mean and the z score. Using this table will indicate whether the calculated z score is within the realm of chance or if the z score is so different from the mean that the sample mean is unlikely to have happened by chance.

Example 7.23: A large distributor of cosmetics has kept his outstanding accounts receivable to a mean time of 18 days over the past year. This average is considered a standard by which to measure the efficiency of the credit and collections department. Management wishes to check if receivables in the current month is over standard and will do this at a significance level of 0.05. A random sample of 100 accounts yields an average of 20 days with a standard deviation of 9 days. What should management conclude?

Solution:

The null hypothesis

$H_0: \mu = 18$ days i.e. the mean time of outstanding accounts is 18 days.

against the **alternative hypothesis**

$H_1: \mu \neq 18$ i.e. the mean time of outstanding accounts is not equal to 18 days.

$\mu_0 = 18, \bar{x} = 20, \sigma = 9$ and $n = 100$

Using the test statistic for large sample test for testing hypotheses concerning a population mean, we find

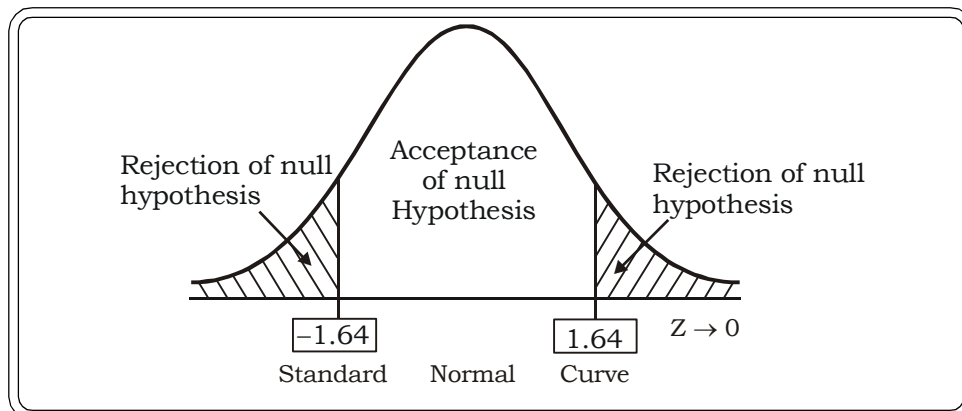
The test statistic

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{20 - 18}{\frac{9}{10}} = 2.22$$

From the z-tables we find the z value at 0.05 level of significance to be 1.645

Decision

Since the calculated value is greater than the tabulated value we reject the null hypothesis.



Conclusion

At significance level 0.05 we may conclude that sample supports the claim that the mean time of outstanding accounts is not equal to 18 days.

Example 7.24: Suppose we want to know if only (single) children have on an average higher cholesterol levels than the national average. It is known that the mean cholesterol level for all Indians is 190 with a standard deviation of 15. Construct the relevant hypothesis test:

Solution:

The null hypothesis –

$H_0: \mu = 190$ i.e. mean cholesterol level of children is 190

The alternative hypothesis –

$H_1: \mu > 190$ i.e. the mean cholesterol level of children is higher than the national average of 190. In a sample of 100 single children we find that the sample average is 198 and the sample standard deviation is 15.

$$\bar{x} = 198 \text{ and } \sigma = 15.$$

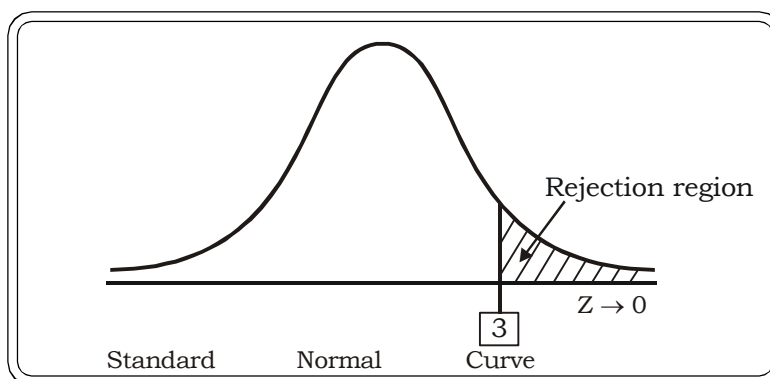
The test statistic

We first calculate the test statistic

$$z = \frac{198 - 190}{15/\sqrt{100}} = 5.33$$

Decision:

This is a right tailed test. Since z is so high (greater than tabled value 3), the probability that H_0 is true is so small that we decide to reject H_0 and accept H_1 .



Conclusion

Therefore, we can conclude that there is substantial statistical evidence indicating single children have a higher cholesterol level on an average than the national average.

Example 7.25: 50 smokers were questioned about the number of hours they sleep each day. We want to test the hypothesis that the smokers need less sleep than the general public, which needs an average of 7.7 hours of sleep.

A. Compute a rejection region for a significance level of 0.05.

B. If the sample mean is 7.5 and the standard deviation is 0.5, what can you conclude?

Solution

First, we write down the null and alternative hypothesis :

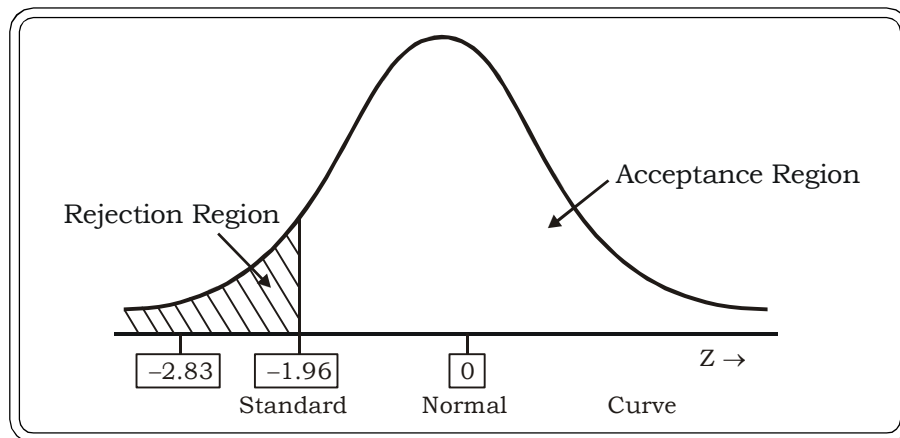
$H_0: \mu = 7.7$ i.e. smokers need to sleep for an average of 7.7 hours

$H_1: \mu < 7.7$ i.e. smokers need less sleep than the general requirement of 7.7 hours

This is a left tailed test. The z score that corresponds to 0.05 is -1.96 . The critical region is the area that lies to the left of -1.96 . If the z-value is less than -1.96 , then we will reject the null hypothesis and accept the alternative hypothesis. If it is greater than -1.96 , we will fail to reject the null hypothesis and say that the test was not statistically significant.

We have the following test statistic

$$z = \frac{7.5 - 7.7}{0.5/\sqrt{50}} = -2.83$$

**Decision:**

Since -2.83 is to the left of -1.96 , it is in the critical region.

Conclusion

Hence we reject the null hypothesis and accept the alternative hypothesis. We may conclude that perhaps smokers need less sleep than the general average of 7.7 hours.

Calculation of Tabled Value from Z Table

1. Select the level of significance.
2. Select the tail: One tail or two tail

Example: Two Tail (5% level of significance)

Total Area

Acceptance $1 - 0.05 = 0.95$ [Total area of Normal Curve is 1], Rejection = 0.05

For each side

Acceptance Region = $0.95/2 = 0.475$, Rejection Region = $0.05/2 = 0.025$

3. Locate from the Z table where 0.4750 of area falls.

Ans. It is 1.96

When exact value is not given in The Z table, use interpolation

Example: 1% level of Significance (Two tailed).

Acceptance region for one side of Normal Curve is $(0.5 - .005 = 0.4950)$

Ordinate	Area
2.57	0.4949
	0.4950
2.58	0.4951

For difference of .0002(.4951-.4949) in area the difference in ordinate is .01(2.58-2.57).

For difference of .0001 in area the difference in ordinate will be $(.01/.002 \times .0001 = .005)$

Thus the tabled value at 1% level of significance is $2.57 + .005 = 2.58$ for two tailed Z test.

7.3.4.2 One Sample t Test for Mean (when population variance is unknown)

In case of small samples (< 30 on an average) and also when the population standard deviation is not known, the t- test for mean is used. The t-test was developed by W. S. Gossett, a statistician employed at the Guinness brewery. However, because the brewery did not allow employees to publish their research, Gossett's work on the t-test appears under the name "Student" (and the t-test is sometimes referred to as "Student's t-test.") Gossett was a chemist and was responsible for developing procedures for ensuring the similarity of batches of Guinness. The t-test was developed as a way of measuring how closely the yeast content of a particular batch of beer corresponded to the brewery's standard. This statistic has been discussed in the chapter on sampling distributions.

One of the advantages of the t-test is that it can be applied to a relatively small number of cases. It was specifically designed to evaluate statistical differences for samples of 30 or less.

A one sample t-test is a hypothesis test for answering questions about the mean where the data are a random sample of independent observations from an underlying normal distribution with mean μ and variance σ^2 :

The null hypothesis for the one sample t-test is:

$$H_0: \mu = \mu_0 \text{ (where } \mu_0 \text{ known)}$$

That is, the sample has been drawn from a population of given mean and unknown variance (which therefore has to be estimated from the sample).

This null hypothesis, H_0 is tested against any one of the following alternative hypotheses, depending on the question posed:

$$H_1 : \mu \neq \mu_0 ,$$

$$H_1 : \mu > \mu_0 \quad \text{or}$$

$$H_1 : \mu < \mu_0$$

The first hypothesis is a two-tailed hypothesis.

The second one is a right tailed hypothesis with the rejection region lying entirely on the right side of the normal curve.

The next one is a left tailed hypothesis with the rejection region lying entirely on the left side of the normal curve. Here, the test statistic is defined as:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

where

\bar{x} = sample mean

μ = population mean under null hypothesis

s = sample standard deviation = $\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

n = sample size

$t \sim t(n-1)$

After calculating the t score by using the above statistic, it is compared with the critical t score from the table at a given level of significance and $(n-1)$ degrees of freedom.

Decision Rule:

The decision rule is to reject H_0 if the calculated value of t lies in the rejection region and accept otherwise.

Example: 7.26: It is required to test whether the temperature required to damage a computer on an average is less than 110 degrees. Because of the price of testing, a sample of twenty computers was tested to see what minimum temperature would damage the computer. It was observed that the damaging temperature averaged 109 degrees with a standard deviation of 3 degrees. Use 0.01 level of significance to test if the damaging temperature is less than 110°.

Solution:

We test the *null hypothesis*

$H_0: \mu = 110$ i.e the average temperature which causes damage to a computer is 110°.

$H_1: \mu < 110$ i.e the average temperature which causes damage to a computer is less than 110°.

The sample size is 20.

We compute *the test statistic*:

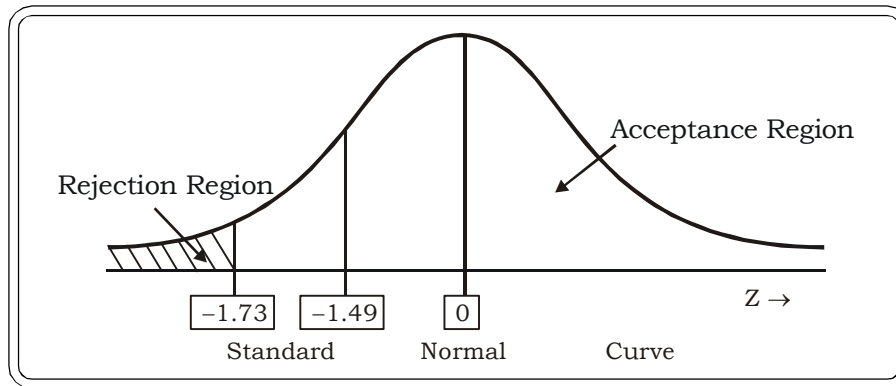
$$t = \frac{109 - 110}{3/\sqrt{20}} = -1.49$$

This is a left tailed test, so we can refer to the t-table with 19 degrees of freedom at 5% level of significance to obtain the tabulated value of t which is

$$t_{19} = -1.73$$

Decision:

Since $-1.49 > -1.73$



We see that the test statistic does not fall in the critical region. We fail to reject the null hypothesis.

Conclusion

We conclude that there is insufficient evidence to suggest that the temperature level which causes damage to a computer on an average less is than 110 degrees.

Example: 7.27: The 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, 68. Test whether the mean price of the share in the month is more than 65.

Solution:

The null hypothesis-

$H_0: \mu = 65$ i.e. the mean price of the share in a month is Rs 65.

The alternative hypothesis

$H_1: \mu > 65$ i.e. the mean price of a share in a month is more than Rs 65.

The test statistic:

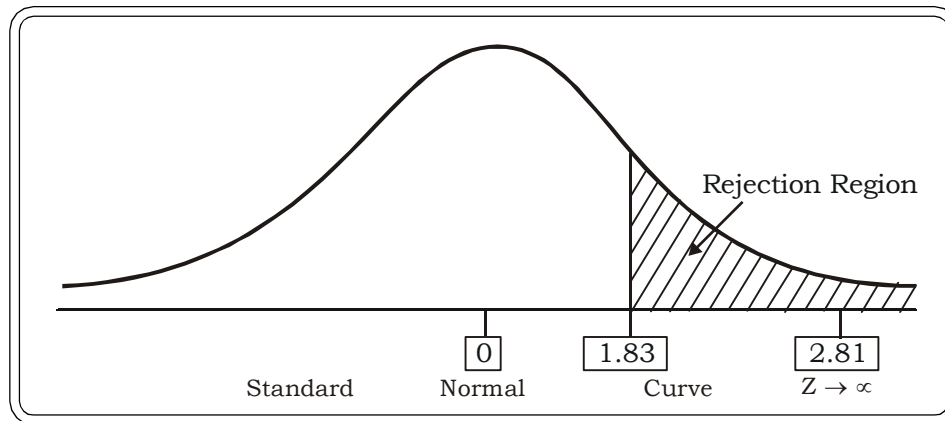
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Here $\bar{x} = \frac{\sum x}{n} = 67.5$

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = 2.81$$

Thus calculated t: $t = \frac{67.5 - 65}{2.8/\sqrt{10}} = \frac{2.5}{0.89} = 2.81$

And the tabulated t: $t_{9,05} = 1.833$ (from t-table)

**Decision:**

The Null hypothesis is rejected since calculated value of t is greater than the tabulated value of t and falls in the rejection region. And the alternative hypothesis may be accepted.

Conclusion

There is sufficient statistical evidence to conclude that the mean price of the share in a month is more than Rs 65.

Example: 7.28: A restaurant near a railway station has been having average sales of 500 cups of tea per day. Currently, because of the development of a bus stand nearby, the owner expects to increase his average per day sales. During the first 12 days after the inauguration of the bus stand the daily sales were observed to be:

550, 570, 490, 615, 505, 580, 570, 460, 600, 580, 530, 526

On the basis of this sample information can one conclude that the restaurant's sales have increased at 5% level of significance?

Solution:

The null and alternative hypotheses are as follows:

$H_0: \mu = 500$ i.e. the average sales is 500 cups per day.

$H_1: \mu > 500$ i.e. the average sales is more than 500 cups per day.

As the sample size is small and the population standard deviation is not known, the t test is applicable here.

The test statistic:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

To calculate out \bar{x} and s the following computations are required

S.No.	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	550	2	4
2	570	22	484
3	490	-58	3364
4	615	67	4489
5	505	-43	1849
6	580	32	1024
7	570	22	484
8	460	-88	7744
9	600	52	2704
10	580	32	1024
11	530	-18	324
12	526	-22	484
$N = 12$	$\sum X_i = 6576$		$(x_i - \bar{x})^2 = 23978$

The test statistic:

We first calculate \bar{x} and s as follows,

$$\bar{x} = \frac{\sum x}{n} = \frac{6576}{12} = 548$$

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{23978}{12-1}} = 46.68$$

$$\text{Calculated } t: t = \frac{548 - 500}{46.68/\sqrt{12}} = \frac{48}{13.49} = 3.558$$

The tabled value of t at $12 - 1 = 11$ degree of freedom and 5% level of significance is 1.796.

Decision:

The calculated t value is 3.558 and it lies in the rejection region. So, the null hypothesis is rejected.

Conclusion

So, we may conclude that the sales of the restaurant seems to have increased, with the opening up of the bus stand nearby.

Example 7.29: An automobile tyre manufacturer claims that the mean number of trouble free kilometers given by a new tyre is 36,000 kms.

In a random sample of 25 randomly picked tyres, the mean mileage is 38,900 kms with a sample standard deviation of 9000 kms. At 5% level of significance, test the manufacturer's claim.

Solution:

The null and the alternative hypothesis

$H_0 : \mu = 36,000$ i.e. mean mileage of the new tyre is 36,000 kms.

$H_1 : \mu \neq 36,000$ i.e. the mean mileage of new tyre is not 36,000 kms.

$n = 25$

$s = 9000$ kms.

$\bar{x} = 38,900$ kms.

The test statistic

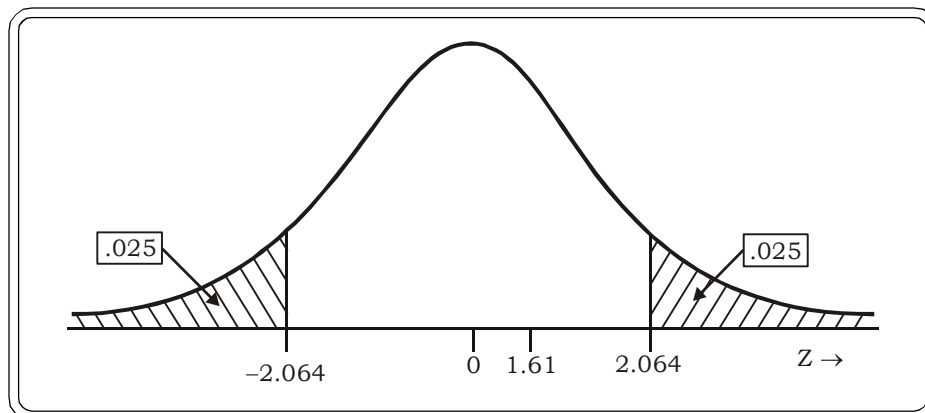
$$\begin{aligned} t &= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1} \\ &= \frac{38,900 - 36,000}{\frac{9000}{5}} \\ &= \frac{2900}{1800} \\ &= 1.61 \end{aligned}$$

We now compare this value with the table value of $t_{0.025,24}$

(Since it is a two tailed test, $\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$)

$t_{0.025,24} = 2.064$

Decision:



Since calculate $t <$ tabulated t at 5% level of significance, we may accept the null hypothesis.

Conclusion

At 5% level of significance, we may conclude that the manufacturers claim seems to be quite valid.

7.3.4.3 One Sample Z Test for Proportion

This section deals with the treatment of data of a qualitative nature. For example, defectives or non-defectives in a sample of parts. Our objective will be to test a hypothesis regarding the population proportion of the attribute being studied. The null hypothesis in this case is of the form,

$H_0: P = P_0$ i.e. population proportion is P_0 .

where P_0 lies between 0 and 1

against any one of the alternatives

$H_1 : P > P_0$ (Right tailed test)

$H_1 : P < P_0$ (Left tailed test)

$H_1 : P \neq P_0$ (Two tailed test)

The test statistic is,

$$z = \frac{\frac{x}{n} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} \sim N(0, 1)$$

where x – number of units possessing the attribute

n – sample size

x/n – sample proportion of units possessing the attribute.

The decision is to reject the null hypothesis if calculated z is greater than the tabulated z .

Example: 7.30: The sponsors of a fashion show at the trade fair believes that the audience is equally divided between male and females. Out of 300 persons attending the fair per day there were 170 males. Test how far the sponsors are correct at 5% level of significance.

Solution:

Let p = sample proportion of males = $170/300 = 0.57$

$n = 300$

Null & Alternative Hypothesis

$H_0 : P_0 = \frac{1}{2}$ i.e. the proportion of males in the audience is half.

$H_1 : P_0 \neq \frac{1}{2}$ i.e. the proportion of males and females is not equal.

Test statistic

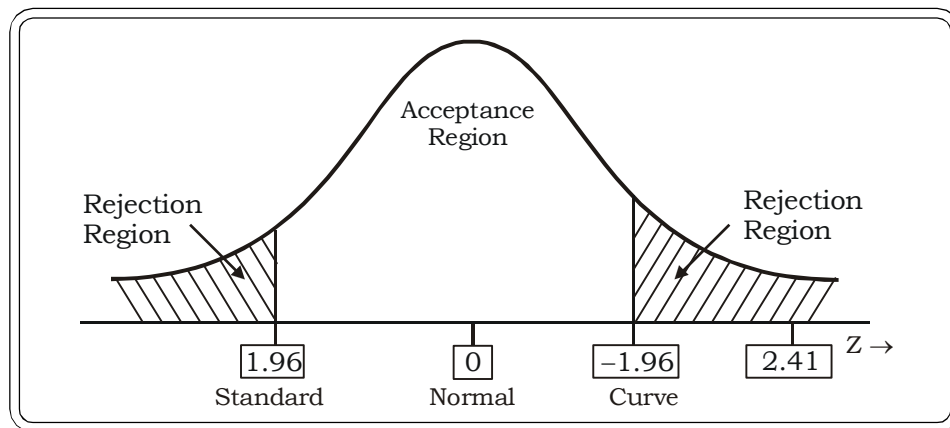
$$z = \frac{\frac{x}{n} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$$

Here $x = 170$, $n = 300$

$$P_0 = \frac{1}{2}$$

Thus

$$z = \frac{0.57 - 0.5}{\sqrt{\frac{0.5(0.5)}{300}}} = \frac{0.07}{0.029} = 2.41$$



Tabulated $z = 1.96$

Decision

Since the calculated value of z is higher than the tabulated value of z , we may reject the null hypothesis and accept the alternative hypothesis.

Conclusion

Thus, the sponsors are not correct in assuming that their audience consists of men & women in equal proportion.

Example: 7.31: A machine is known to produce 20% defective screws. After the machine was repaired, it was found that it produced 25 defective screws in the first run of 100. Evaluate if it is true that after the repairs the proportion of defective screws has been reduced. (Use $\alpha = 0.01$).

Solution:

Let P = Historic proportion of defective screws

$$x = 25, n = 100$$

The null hypothesis

$H_0: P = 0.2$ i.e. the proportion of defectives is 0.2

The alternative hypothesis

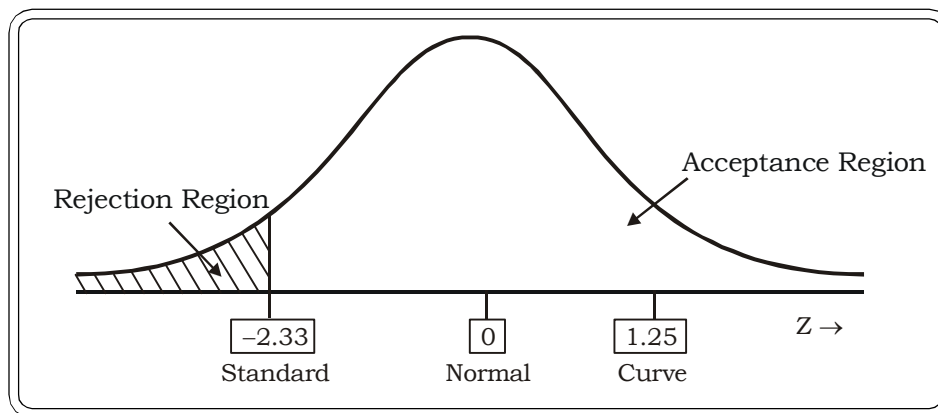
$H_1: P < 0.2$ i.e. the proportion of defectives is less than 0.2

The test statistic:

$$z = \frac{\frac{x}{n} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{\frac{25}{100} - 0.2}{\sqrt{\frac{(0.2)(0.8)}{100}}}$$

$$= 1.25$$

Tabulated value: $z = -2.33$



Decision:

We may accept H_0 , since calculated value is much greater than the tabulated value and it lies in the acceptance region of the curve.

Conclusion

There is no statistical evidence to reject H_0 . So we may say that the proportion of defective screws has not been reduced after the repairs.

7.3.5 Two Sample test

In two sample tests, we test hypothesis regarding the difference of means of two populations and the difference of proportions of an attribute in two different populations.

7.3.5.1 Two-Sample Z Test for Difference of Two Means

Consider two populations with means μ_1 and μ_2 respectively. Suppose our interest lies in testing if there is any difference in the two populations with respect to their means.

The null hypothesis in this case is

$$H_0: \mu_1 - \mu_2 = 0$$

i.e. there is no difference in the population means, against any one of the alternative hypothesis.

$H_1: \mu_1 \neq \mu_2$ i.e. there is significant difference between the two populations.

or $H_1: \mu_1 > \mu_2$ (right tailed test) i.e. the mean of the 1st population is greater than the mean of the second population

or $H_1: \mu_1 < \mu_2$ (left tailed test) i.e. the mean of the 1st population is less than the mean of the second population

To test the null hypothesis against any one of the alternative hypothesis, we now consider two random samples of size m and n drawn from the two respective populations. The sample means are then calculated and let

\bar{x}_1 = mean of the sample drawn from the 1st population.

\bar{x}_2 = mean of the sample drawn from the 2nd population

The population variances σ_1^2 and σ_2^2 are assumed to be known

The test statistic is:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0,1)$$

\bar{x}_1 = mean of the sample drawn from the first population

\bar{x}_2 = mean of the sample drawn from the second population

σ_1^2 = variance of the first population.

σ_2^2 = variance of the second population.

m = size of sample from population 1.

n = size of the sample from population 2.

When variances are known and equal which is the usual assumption, we replace σ_1^2 and σ_2^2 by σ^2 in the test statistic.

The rest of the testing procedure remains the same.

Example 7.32: A factory conducted an experiment to compare the productivity of two machines – machine A and machine B.

Machine A was kept under observation for 40 hrs and machine B kept under observation for 50 hrs. The average productivity of units produced per hour is 62.5 for machine A and 60 for machine B. Historically, the standard deviation of the two machines are 3.2 and 2.9 respectively. At 0.10 level of significance, test whether productivity of machine A is better than productivity of machine B.

Solution:

The null hypothesis

$H_0: \mu_1 = \mu_2$ i.e. there is no significant difference in the productivity of machine A & B.

The alternative hypothesis

$H_1 : \mu_1 > \mu_2$ i.e. productivity on machine A is better than the productivity on machine B.

Given: $\bar{x}_1 = 62$; $\bar{x}_2 = 60$; $\sigma_1 = 3.2$; $\sigma_2 = 2.9$; $m = 40$; $n = 50$

The test statistic

$$Z = \frac{62 - 60}{\sqrt{\frac{(3.2)^2}{40} + \frac{(2.9)^2}{50}}} = \frac{2}{\sqrt{\frac{10.25}{40} + \frac{8.41}{50}}} = \frac{2}{\sqrt{0.26 + 0.17}} = \frac{2}{0.66} = 3.03$$

Tab $Z_{0.05} = 1.64$

Decision:

Since calculated Z is greater than the tabulated Z , we may reject the null hypothesis and accept the alternative hypothesis at 5% level of significance.

Conclusion

There is enough statistical evidence to indicate that productivity of machine A is more than the productivity of machine B.

7.3.5.2 Two Sample t-Test for Difference of Two Means

A two-sample z-test is a hypothesis test for answering questions about the mean when the data are collected from two random samples of independent observations, each from an underlying normal distribution i.e.:

$$N(\mu_i, \sigma_i^2) \text{ where } i = 1, 2$$

A t test is preferred over the z test when the sample size is small (usually <30) and the population variances are unknown.

While carrying out a two-sample t-test, it is usual to assume that the variances for the two populations are equal, though unknown, that is:

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

The *null hypothesis* for the two-sample t-test is:

$$H_0 : \mu_1 = \mu_2$$

the two samples have both been drawn from populations having the same mean.

This null hypothesis is tested against any one of the following *alternative hypotheses*, depending upon the question posed.

$H_1 : \mu_1 \neq \mu_2$ (Two - tailed test)

$H_1 : \mu_1 > \mu_2$ (Right - tailed test)

$H_1 : \mu_1 < \mu_2$ (Left - tailed test)

The next step is the selection of the test statistic.

When population variances are unknown and equal and the sample sizes are small we calculate a pooled variance using the sample data and the formula:

$$s_p^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}$$

The test statistic to be used in this case is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

The statistic follows the t-distribution with $(m + n - 2)$ degrees of freedom.

The decision rule:

The decision rule is to reject the null hypothesis if the calculated statistic is larger than the tabulated statistic, at $m + n - 2$ degrees of freedom and the required level of significance.

Provided m and n are sufficiently large. We use the following statistic, which follows a standard normal distribution

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} \sim N(0, 1)$$

Example: 7.33: A company has drawn sample of sales of a new product in two shops, located in two different cities for their new launched product. Data presented are as follows:

City	Mean sales	Standard deviation	Sample size
Delhi	89	2.55	7
Kolkata	82	2.37	4

The company would like to assess the difference in sales in two cities. Find out at 5% level of significance, if there is any difference between the mean sales of the new product in the two cities.

Solution:

Consider the **null hypothesis** that the means of two populations are equal. Thus we can write

$H_0 : \mu_1 = \mu_2$ i.e. the mean sales of the new product in both two cities are same.

and the **two tailed alternative** hypothesis as

$H_1 : \mu_1 \neq \mu_2$ i.e. the mean sales of the new product in Delhi & Kolkata are different..

The test statistic:

As in the given problem, the population variance is not given and the sample size is small we use the t test of difference between two means. We first calculate the pooled variance.

$$s_p^2 = \frac{6 \times 2.55^2 + 3 \times 2.37^2}{7 + 4 - 2}$$

$$= \frac{39.01 + 16.85}{9} = \frac{55.86}{9} = 6.21$$

The test statistic is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}$$

$$= \frac{89 - 82}{\sqrt{6.21} \sqrt{\frac{1}{7} + \frac{1}{4}}}$$

$$= \frac{7}{2.49 \sqrt{0.14 + 0.25}} = \frac{7}{1.56}$$

$$= 4.49$$

In this case, the tabled value of t at $7 + 4 - 2 = 9$ degrees of freedom and 5% level of significance is 2.26.

Decision

Reject H_0 (at 5% level of significance) since calculated value lies in the rejection region.

Conclusion

Thus, it can be concluded that the mean sales of the new launched product of the company is not same both in Delhi & Kolkata.

Example 7.34: Two business schools want to determine if the mean scores of CAT for the students in their institutes are similar. A simple random sample of 20 students is taken from the first business school (A) and a simple random sample of 25 students is taken from the second business school (B). The survey yields the following results:

The average score of School A = 750

The average score of School B = 650

The sample standard deviation of school A: 80

The sample standard deviation of school B: 90

Test if there is significant difference in the mean CAT scores of the students of school A & school B (use 0.01 level of significance).

Solution:

Since sample standard deviations are known and population standard deviations are unknown we can apply the t - test for difference of 2 means. However, since sample sizes are reasonably large, the normal distribution would also be appropriate.

The null hypothesis

$H_0 : \mu_1 = \mu_2$ i.e. there is no significant difference in the mean CAT scores of the students of the two business schools.

The alternative hypothesis

$H_0 : \mu_1 \neq \mu_2$ i.e. there is significant difference in the mean CAT scores of the students of the two business schools.

Given:

$$\bar{x}_1 = 750; \bar{x}_2 = 650; s_1 = 80; s_2 = 90$$

$$n_1 = 20; n_2 = 25$$

The test statistic

$$\begin{aligned} Z &= \frac{750 - 650}{\sqrt{\frac{80^2}{20} + \frac{90^2}{25}}} \\ &= \frac{100}{\sqrt{320 + 324}} \\ &= \frac{100}{\sqrt{644}} = \frac{100}{25.38} \\ &= 3.94 \end{aligned}$$

The tabulated value of Z at 0.05 level of significance is 1.96.

Decision:

There is not enough statistical evidence to accept the null hypothesis. So we may accept the alternative hypothesis.

Conclusion

There seems to be significant difference in the mean CAT scores of the students in the two business schools.

Example 7.35: In 16 half-hour evening programs, the mean time devoted to commercials was 6.4 minutes with $s_1 = 2$ mins. In 16 half hour morning programs the mean time was 5.8 minutes with $s_2 = 1.5$ minutes. Test, at 10% level of significance, if the data indicates that the mean time devoted to commercials is significantly less in the morning? (Assume that the populations are normally distributed with same but unknown variances)

Solution:**The null and the alternative hypothesis**

$H_0 : \mu_1 = \mu_2$ i.e. mean time devoted to commercials in the morning and evening is same.

$H_1 : \mu_1 < \mu_2$ i.e. mean time devoted to commercials is less in the morning as compared to evening.

This is a left tailed test.

Morning commercial sample information:

$$\bar{x}_1 = 5.8 \text{ minutes}$$

$$s_1 = 1.5 \text{ minutes}$$

$$n_2 = 16$$

Evening commercial sample information:

$$\bar{x}_1 = 6.4 \text{ minutes}$$

$$s_1 = 2 \text{ minutes}$$

$$n_2 = 16$$

The test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{where } s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

(since population variances are unknown but equal)

$$\therefore s^2 = \frac{15 \times (1.5)^2 + 15 \times 2^2}{15 + 15 - 2}$$

$$= \frac{33.75 + 60}{28}$$

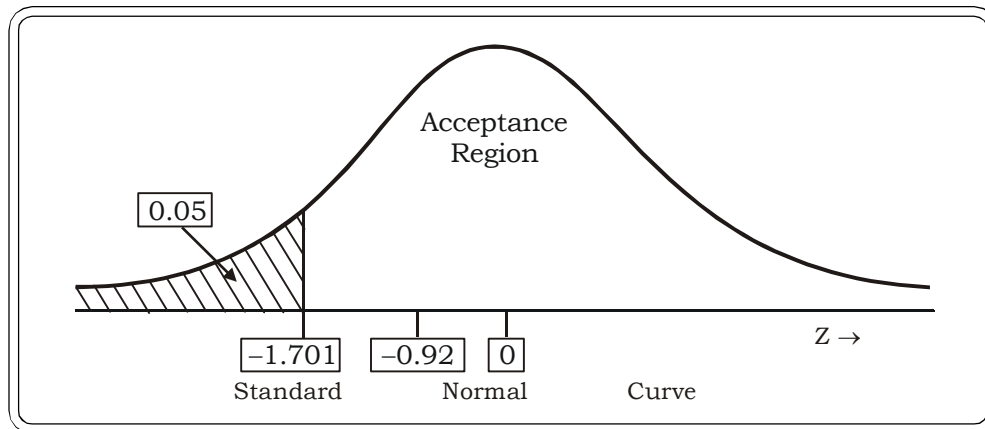
$$= 3.35$$

$$\Rightarrow s = 1.83$$

$$\therefore t = \frac{5.8 - 6.4}{1.83 \sqrt{\frac{2}{16}}}$$

$$= \frac{-0.6}{0.65}$$

$$= -0.92$$



Degree of freedom = 28

Tabulated $t_{0.05, 28} = -1.701$

Decision

Since calculated t lies in the acceptance region, we may accept the null hypothesis at 5% level of significance.

Conclusion

We may conclude that the mean time devoted to commercials in the morning and evening are not significantly different.

7.3.5.3 Paired t Test (for correlated or dependent samples)

The two sample paired t test is used to test the difference of two population means when the two samples are correlated i.e. there exists a one -to- one correspondence between the values of the sample. Usually, measurements are related to the same subject before and after a process change. For example, measurements could be related to blood pressure levels before and after the administration of a drug to test the efficacy of the drug.

Let $X_{11}, X_{12}, X_{13} \dots \dots X_{1n}$ represent n observations from the first sample and

Let $X_{21}, X_{22}, X_{23} \dots \dots X_{2n}$ represent n observations from the corresponding second sample.

Then $d_1 = X_{11} - X_{21}, d_2 = X_{12} - X_{22}, \dots \dots \dots, d_n = X_{1n} - X_{2n}$ are the difference scores between the corresponding observations of the two samples.

These scores are regarded as observations from a single sample. And assuming that these are randomly and independently drawn from a population that is normally distributed, the paired t-test can now be applied to test for significant difference in the population means.

The null hypothesis to be tested is that there is no significant difference in the means of the two related samples i.e.

$$H_0 : \mu_1 = \mu_2$$

against the alternative hypothesis that there is significant difference in the two means i.e.

$$H_1 : \mu_1 \neq \mu_2$$

The **test statistic** is

$$t = \frac{\bar{d}}{s/\sqrt{n}} \sim t_{n-1}$$

where $\bar{d} = \frac{\sum_{i=1}^n d_i}{n}$

and $s = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}}$

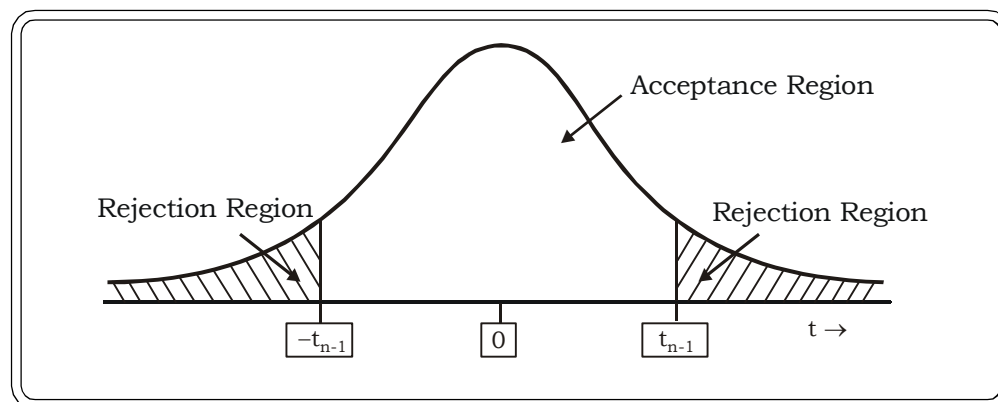


Figure 7.11

Critical Region for Paired t-test

Decision rule

Reject H_0 if $t > t_{(n-1)}$ or $t < -t_{(n-1)}$ at $\alpha\%$ level of significance.

Else we may accept H_0 .

Example 7.36: The Peak Expiratory Flow Rate (PEFR) of 9 asthma patients was taken before and after a walk on an extremely cold winter day for comparing the rates. The following data was obtained:

Patient	Before	After
1	312	300
2	242	201
3	340	232
4	388	312
5	296	220
6	254	256
7	391	328
8	402	330
9	290	231

Test whether there is any significant difference between the PEFR of asthma patients before and after a walk on a cold winter day.

Solution:

The null hypothesis

Ho: there is no significant difference of the PEFR of the asthma patients before and after a walk on a cold winter day.

The alternative hypothesis

H₁: there is significant difference of the PEFR of the asthma patients before and after a walk on a cold winter day

Calculations: We first make the following table:

Patient No.	Before (x_1)	After (x_2)	$d_i = x_1 - x_2$
1	312	300	12
2	242	201	41
3	340	232	108
4	388	312	76
5	296	220	76
6	254	256	-2
7	391	328	63
8	402	330	72
9	290	231	59
			$\Sigma d_i = 505$

$$n = 9$$

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} = 56.11$$

$$s = 11.39$$

The test statistic

$$t = 14.80$$

$$n-1 = 8$$

Tabulated $t_8 = 2.306$ at $\alpha = 0.05$

Decision rule

There is not enough evidence to accept the null hypothesis, as the calculated value is much greater than the tabulated value.

Conclusion

There seems to be significant difference in the effect of the cold winter day on the PEFR of the asthma patients.

Example 7.37: In a bid to prepare the city for an international games event in 2010, the city Government has decided to train the taxi drivers of the city to converse in English and increase awareness of the city drivers. 5 were given intense training on English in various areas of city. A test was taken before commencement of the training and after commencement of the training to determine how effective the training program has been. The data below lists the scores of the five individuals before and after the training.

Number	Score Before Training	Score After Training
1	3	7
2	2	6
3	4	7
4	3	5
5	5	7

At 0.05 level of significance, has the training program been successful?

Solution:

The null hypothesis and the alternative hypothesis:

$H_0 : \bar{d} = 0$ i.e. there is no significant difference in the scores before and after training.

$H_1 : \bar{d} > 0$ i.e. there is significant improvement in the scores of the drivers after the training.

This is a one – tailed test.

Calculations:

Before	After	d	(d - \bar{d}) ²
7	3	+4	1
6	2	+4	1
7	4	+3	0
5	3	+2	1
7	5	+2	1

$$\bar{d} = \frac{\sum_{i=1}^5 d_i}{5} = \frac{15}{5} = 3$$

$$s_d = \sqrt{\frac{\sum_{i=1}^5 (d_i - \bar{d})^2}{5-1}} = \sqrt{\frac{4}{4}} = 1$$

The test statistic

$$t = \frac{\frac{\bar{d}}{s_d}}{\frac{1}{\sqrt{n}}} = \frac{\frac{3}{1}}{\frac{1}{\sqrt{5}}} = 6.67$$

The critical value of t at 4 degrees of freedom and level of significance 0.05 is 2.132.

Decision:

Since the calculated value is much higher than the table value there is not enough evidence to accept H_0 .

Conclusion

The training given by the Delhi Government to improve the skills of taxi drivers seem to have been successful, since the scores of the drivers seem to have shown a significant improvement.

7.3.5.4 Two Sample Z Test for Difference of Proportions

It is often required to compare & analyze difference between two populations in terms of a categorical variable or attribute. To test the difference between two proportions obtained from independent samples there are two methods- one is the Z- statistic method, which uses an approximated standard normal distribution. The other is a procedure using a χ^2 (chi-square) test statistic, which will be discussed in chapter 10.

Right now our focus is to analyze differences between two proportions on the basis of two independent samples by using a Z test for the difference between two proportions.

Let P_1 be the proportion of success in the first population.

P_2 be the proportion of success in the second population.

The null hypothesis to be tested in this case is

$H_0 : P_1 = P_2$ i.e. proportion of success in the first population is same as the proportion of success in the second population.

Against any one of the alternative hypothesis,

$H_1 : P_1 > P_2$ (Right Tailed Test) i.e. proportion of success in the first population is greater than the proportion of success in the second population.

or $H_1 : P_1 < P_2$ (Left Tailed Test) i.e. proportion of success in the first population is less than the proportion of success in the second population.

or $H_1 : P_1 \neq P_2$ (Two Tailed Test) i.e. proportion of success in the first population is not equal to the proportion of success in the second population.

The test statistic used is

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{m} + \frac{1}{n}\right)}} \sim N(0,1)$$

where

$$\hat{p} = \frac{x_1 + x_2}{m + n}$$

p_1 = proportion of success in 1st sample.

p_2 = proportion of success in 2nd sample.

m = size of 1st sample.

n = size of 2nd sample.

x_1 = no. of successes in 1st sample.

x_2 = no. of successes in 2nd sample.

\hat{p} = pooled estimate of the population proportion of success.

The test statistic Z follows a standard normal distribution as usual.

Decision rule

For a given level of significance H_0 is rejected if calculated z is higher than the tabulated value and vice versa.

Let us now look at an example, which uses this test.

Example 7.38: A survey conducted for a study of a no frill Airlines called Easy Fly Airlines had the question “will you travel by Easy fly Airlines again?”

163 of 227 men answered, “Yes” and 154 of 262 women answered “No”

At 0.05 level of significance is there evidence of a significant difference in preference of men & women to fly the no frill Easy Fly Airlines again?

Solution:

$H_0 : P_1 = P_2$ i.e. there is no significant difference in the proportion of men & women who prefer to fly Easy fly Airlines.

$H_1 : P_1 \neq P_2$ i.e. there is significant difference in the proportion of men & women who want to fly the no frill airline.

The critical values at 0.05 level of significance are -1.96 & 1.96 .

The Test Statistic

We calculate the sample proportions as follows:

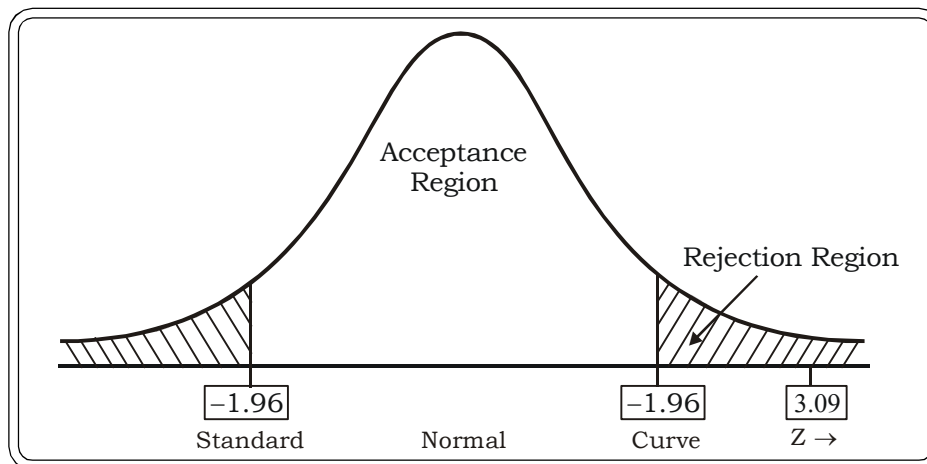
$$p_1 = \frac{163}{227} = 0.718$$

$$p_2 = \frac{154}{262} = 0.585$$

$$p_1 - p_2 = 0.133$$

and the pooled estimate of population proportion of successes

$$\hat{p} = \frac{x_1 + x_2}{m + n} = \frac{317}{489} = 0.648$$



$$\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{m} + \frac{1}{n}\right)} = \sqrt{(0.228)(0.008)} = 0.043$$

$$\text{Then, } Z = \frac{0.133}{0.043} = 3.09$$

Decision Rule

There is insufficient evidence to accept the null hypothesis since the calculated test statistic lies in the rejection region.

Conclusion

Men & women differ significantly in terms of their preference to fly Easy Fly Airlines in the future.

Example 7.39: A sample of 500 shoppers at a mall in NCR Delhi were asked the question “do you enjoy shopping for new CD’s?”

Of 240 males 180 answered, “Yes” & of 260 females 150 answered, “Yes”.

Is there evidence of a significant difference between proportion of males & females in the population who enjoy shopping for the latest music in a mall? (Use 0.01 level of significance).

Solution:

The null hypothesis

H_0 : There is no difference in the proportion of males and females in the population who enjoy shopping for the latest music in a mall.

The alternative hypothesis

H_1 : There is significant difference in the proportion of males and females in the population who enjoy shopping for the latest music in a mall.

$$\begin{aligned}x &= 180 & y &= 150 \\m &= 240 & n &= 260 \\p_1 &= \frac{180}{240} = 0.75 & p_2 &= \frac{150}{260} = 0.58 \\ \hat{p} &= \frac{330}{500} = 0.66\end{aligned}$$

$$\begin{aligned}\text{Test statistic } Z &= \frac{p_1 - p_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{m} + \frac{1}{n}\right)}} \\ &= \frac{0.75 - 0.58}{\sqrt{.66(.34)\left(\frac{1}{240} + \frac{1}{260}\right)}} \\ &= \frac{0.17}{\sqrt{0.2244(0.0042 + 0.0038)}} \\ &= \frac{0.17}{0.042} = 4.05\end{aligned}$$

$$z_{0.005} = 2.58 \quad \left(\frac{\alpha}{2} = \frac{0.01}{2} = 0.005\right)$$

Decision: Reject the null hypothesis since calculated Z is greater than tabulated Z at 1% level of significance.

Conclusion: Thus in conclusion, it may be said that there seems to be is a significant difference between proportion of males & females in the population who enjoy shopping for the latest music in a mall.

Selection between z and t test

The selection between a z test and a t test is sometimes difficult. The following table gives a brief summary of the conditions under which to apply the z test and conditions necessary to apply the t test.

Table 7.3
Selection of z and t test

Sample size	Status of Population Variance	
	Given	Not Given
> 30	z	z
< 30	z	t

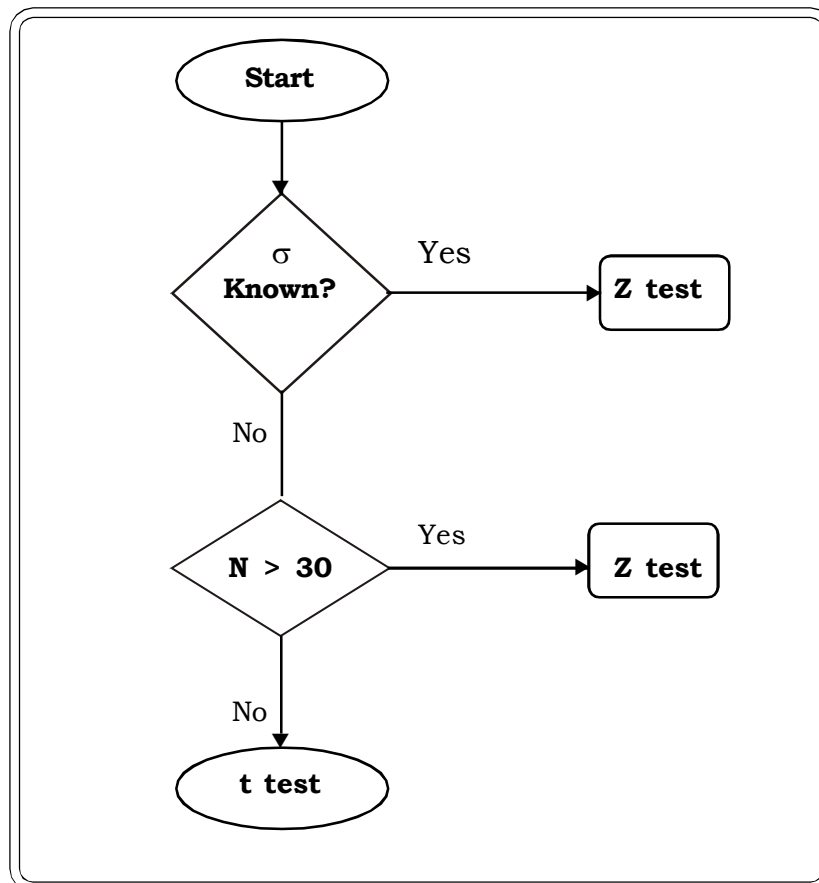


Figure 7.12

Flowchart of selection of z-test and t-test

7.4 CASELETS

Caselet 1: One of the key measures of service quality of any organization is the speed with which it responds to customer complaints. A direct selling organization specializing in vacuum cleaner and water filters had undergone major expansion plans in the past several years. The section handling water filters both for domestic as well as industrial purpose had expanded from 4 installation crews to an installation supervisor, a measurer and 15 installation crews.

During the past year after the expansion was completely put into place there were 15 complaints concerning water filter installation. The following data gives the no. of days between receipt of the complaint and the resolution of the complaint:

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

The installation supervisor claims that the mean number of days between the receipt of the complaint and the resolution of the complaint is 8 days. Is this claim true? What assumption needs to be made in this case?

Caselet 2: Customer Relationship Management (CRM) is a business approach that integrates people, processes & technology to maximize the relations of an organization with all types of customers. CRM in the banking sector is assuming a lot of importance currently. Banks in India like ABN – AMRO, HSBC, ICICI bank, HDFC bank to name a few is actively implementing CRM practices with varied degrees of success. A survey carried out at a particular branch of a bank in the national capital territory of Delhi interviewed employees at various levels for their opinions on implementation of CRM processes in their bank. Out of a total sample of size 40, 56% were male employees & 44% female employees. Two questions posed to the employees were:

Question 1: Does your bank use CRM processes effectively?

Question 2: Do CRM tools in sales benefit you?

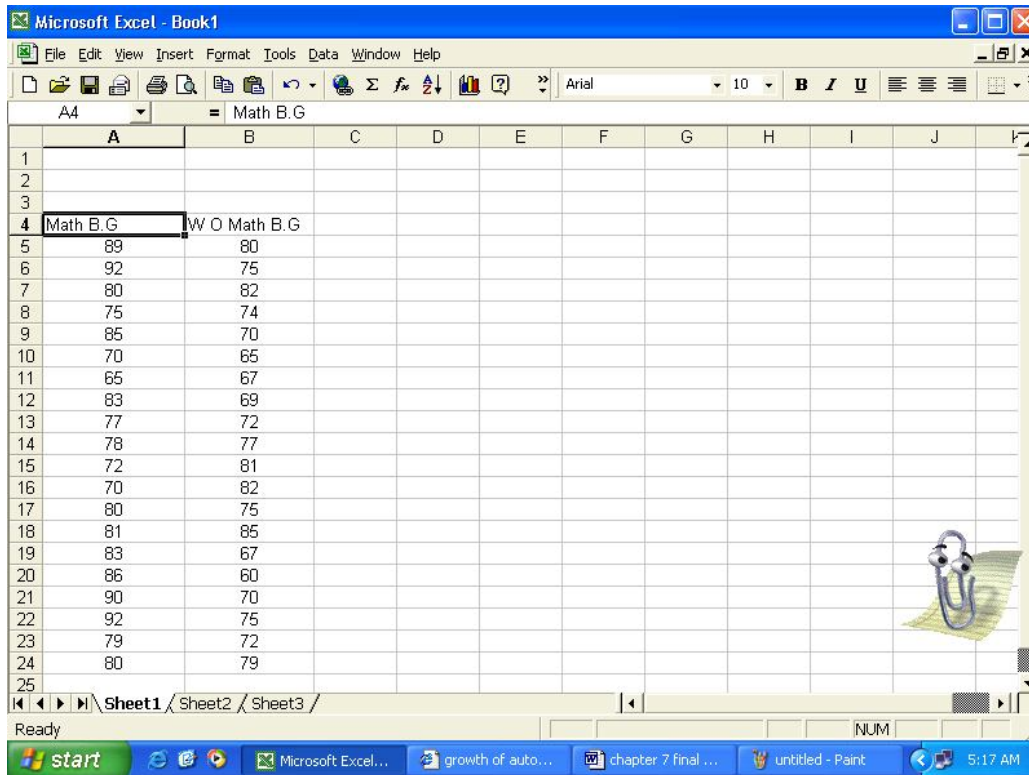
For the 1st Question 45% answered “Yes” and 55% answered no while for the 2nd Question 25 % answered, “Yes” and 75% answered “No”. How best can these data be analyzed? What kind of conclusions will it lead to?

7.5 EXCEL GUIDE

Z – Test for Difference of Two Means (assuming known variances)

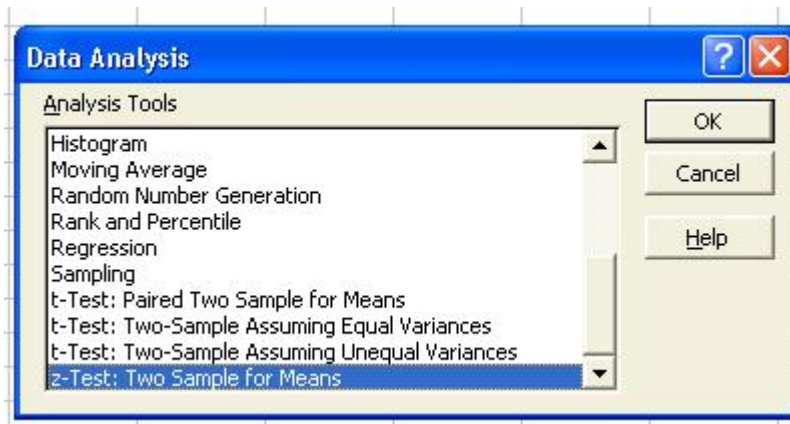
Suppose the marks of students in the quantitative analysis section of CAT examination is compared to see whether students with Math Background have better scores than students without Math Background. The following data gives the marks of students in both cases.

Step 1: Enter the data in a Excel worksheet as follows:



	A	B	C	D	E	F	G	H	I	J
1										
2										
3										
4	Math B.G	W O Math B.G								
5	89	80								
6	92	75								
7	80	82								
8	75	74								
9	85	70								
10	70	65								
11	65	67								
12	83	69								
13	77	72								
14	78	77								
15	72	81								
16	70	82								
17	80	75								
18	81	85								
19	83	67								
20	86	60								
21	90	70								
22	92	75								
23	79	72								
24	80	79								
25										

Step 2: Go to TOOLS & then select DATA ANALYSIS, chose Z – TEST: TWO- SAMPLE for MEANS. Click OK.



Step 3: In Variable 1 Range enter the data of the first variable Cell No. A5 – A24 in this case. In Variable 2 Range enter the data of the second variable Cell No. B5 – B24. In Hypothesized Mean Difference enter 0, since a null hypothesis is that there is no significant difference between the population means. Enter the known variances of the two populations in this case 16 & 25 respectively. In output options chose New Worksheet Ply for the result to be displayed in a new worksheet. Click OK.

z-Test: Two Sample for Means

Input

Variable 1 Range: \$A\$5:\$A\$24

Variable 2 Range: \$B\$5:\$B\$24

Hypothesized Mean Difference: 0

Variable 1 Variance (known): 16

Variable 2 Variance (known): 25

Labels

Alpha: 0.05

Output options

Output Range:

New Worksheet Ply:

New Workbook

OK
Cancel
Help

Step 4: The result is now displayed in a new worksheet. z stat = 4.539797 is the value of the calculated test statistic and z critical tv = 1.959961 is the tabulated t value. The decision & conclusion can be drawn accordingly.

Microsoft Excel - Book1

File Edit View Insert Format Tools Data Window Help

A1 = z-Test: Two Sample for Means

	A	B	C	D	E	F	G	H	I	J	K	L
1	z-Test: Two Sample for Means											
2												
3		Variable 1	Variable 2									
4	Mean	80.35	73.85									
5	Known Var	16	25									
6	Observatio	20	20									
7	Hypothesis:	0										
8	z	4.539797										
9	P(Z<=z) or	2.82E-06										
10	z Critical o	1.644853										
11	P(Z<=z) tv	5.64E-06										
12	z Critical t	1.959961										
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												
23												
24												
25												

Ready Sum=243.3446195 NUM

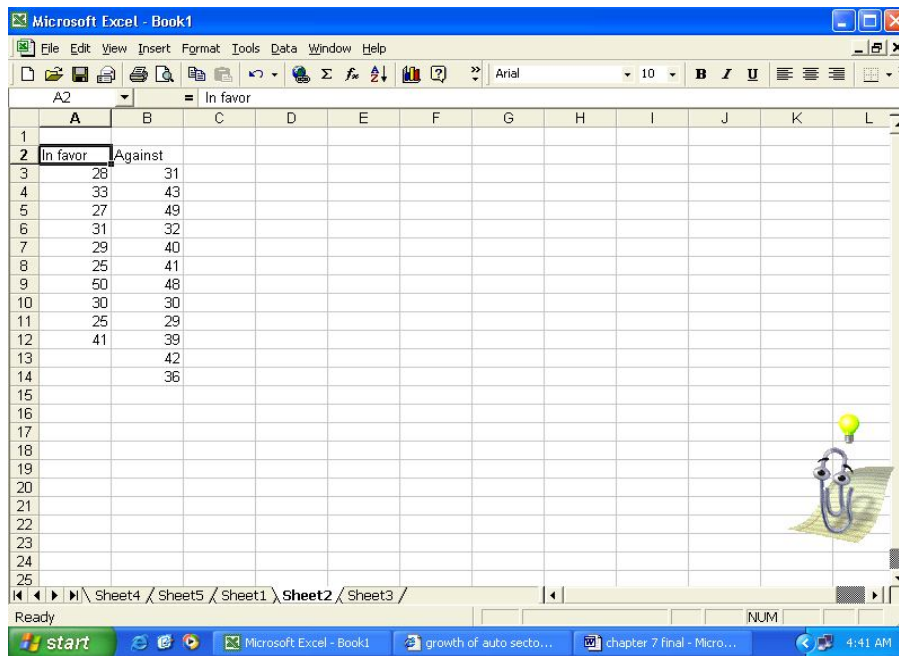
t – Test for Difference of Two Means (assuming equal variances).

Example: Suppose 10 people voted in favor of a co- education hostel in a certain university & 12 voted against co-education hostel. The following data gives the ages of the voters:

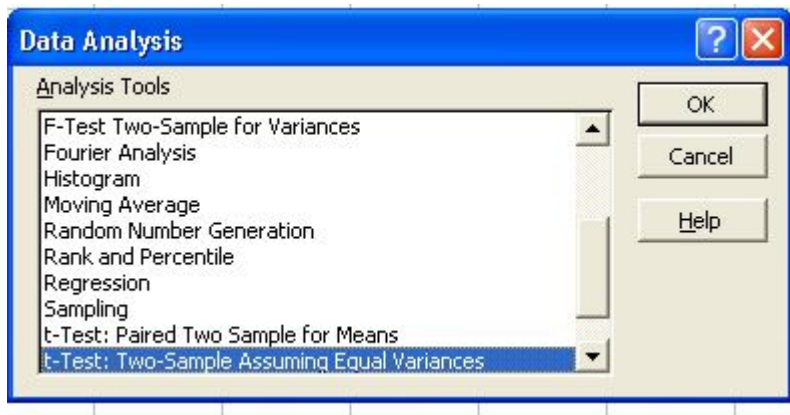
In favor	28	33	27	31	29	25	50	30	25	41		
Against	31	43	49	32	40	41	48	30	29	39	42	36

At 5% level of significance, is the mean age of people voting against the proposition significantly different from the mean age of those voting for it?

Step 1: Enter the data given above in an Excel worksheet as follows:



Step 2: Go to TOOLS & then select DATA ANALYSIS, chose t – TEST: TWO- SAMPLE ASSUMING EQUAL VARIANCES. Click OK.



Step 3: In Variable 1 Range enter the data of the first variable Cell No. A3 - A12 in this case. In Variable 2 Range enter the data of the second variable Cell No. B3 - B14. In Hypothesized Mean Difference enter 0, since a null hypothesis is that there is no significant difference between the population means. In output options chose New Worksheet Ply for the result to be displayed in a new worksheet. Click OK.

t-Test: Two-Sample Assuming Equal Variances

Input

Variable 1 Range:

Variable 2 Range:

Hypothesized Mean Difference:

Labels

Alpha:

Output options

Output Range:

New Worksheet Ply:

New Workbook

OK
Cancel
Help

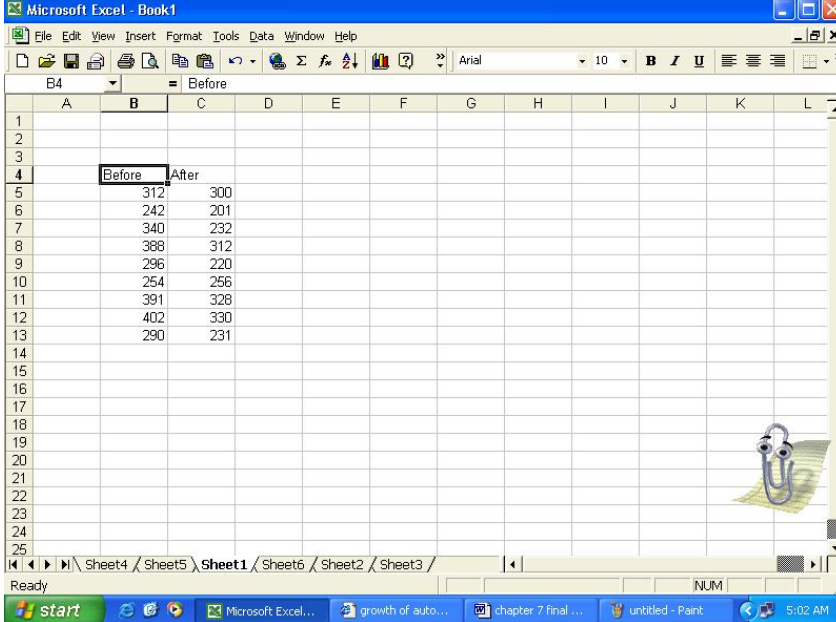
Step 4: The result is now displayed in a new worksheet. t stat = -2.05652 is the value of the calculated test statistic and t critical $t_v = 2.085962$ is the tabulated t value. The decision & conclusion can be drawn accordingly.

	Variable 1	Variable 2	Variable 3
Mean	31.9	38.33333	44.76667
Variance	62.1	46.24242	30.38485
Observatio	10	12	14
Pooled Va	53.37833		53.37833
Hypothesi:	0		0
df	20		20
t Stat	-2.05652		-2.05652
P(T<=t) on	0.026508		0.026508
t Critical oi	1.724718		1.724718
P(T<=t) tw	0.053016		0.053016
t Critical tv	2.085962		2.085962

Paired t - Test for Dependent Samples

This Excel Guide is given for example 7.36

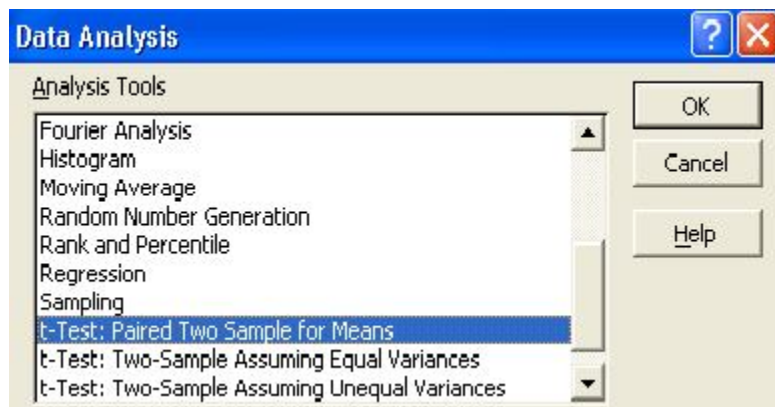
Step 1: Enter the data in a Excel worksheet as follows:



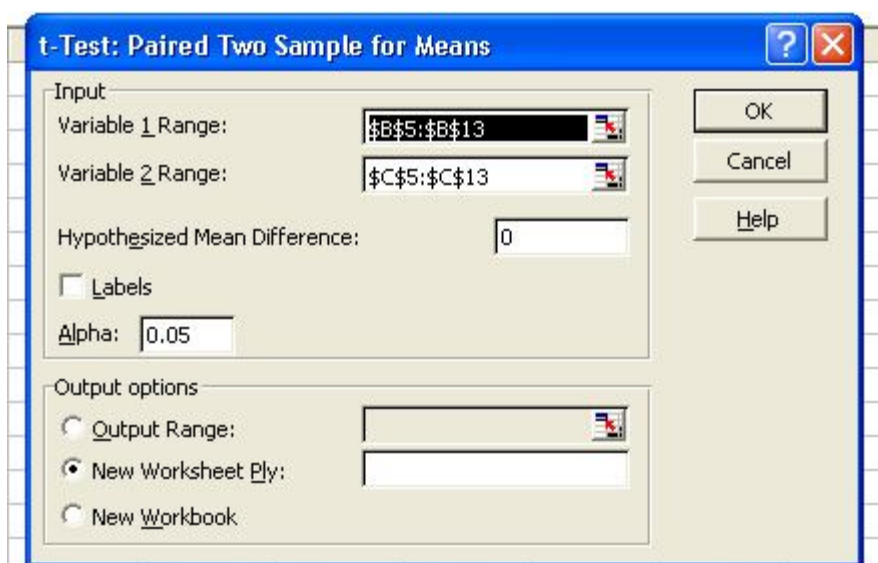
The screenshot shows an Excel worksheet with the following data:

	Before	After
5	312	300
6	242	201
7	340	232
8	388	312
9	296	220
10	254	256
11	391	328
12	402	330
13	290	231

Step 2: Select TOOLS & chose DATA ANALYSIS. Select t – TEST: PAIRED TWO SAMPLE FOR MEANS. Click OK.



Step 3: In Variable 1 Range enter the data of the first variable Cell No. B5 - B13 in this case. In Variable 2 Range enter the data of the second variable Cell No. C5 - C13. In Hypothesized Mean Difference enter 0, since a null hypothesis is that there is no significant difference between the population means. In output options chose New Worksheet Ply for the result to be displayed in a new worksheet. Click OK.



Step 4: The result is now displayed in a new worksheet. t stat = 4.925774 is the value of the calculated test statistic and t critical $t_v = 2.306006$ is the tabulated t value. The decision & conclusion can be drawn accordingly.

	Variable 1	Variable 2
Mean	323.8889	267.7778
Variance	3579.111	2500.694
Observatio	9	9
Pearson C	0.820929	
Hypothesi:	0	
df	8	
t Stat	4.925774	
P(T<=t) on	0.000578	
t Critical o	1.859548	
P(T<=t) tw	0.001156	
t Critical tv	2.306006	

7.6 EXERCISES

- 7.1 What do you mean by estimation? Explain the difference between estimate and estimator.
- 7.2 Define Point and Interval estimates and give examples of each.
- 7.3 What are the properties of a good estimator?
- 7.4 Define Type I and Type II error. Which one is more serious and why?
- 7.5 Explain null hypothesis and alternative hypothesis with suitable example.
- 7.6 Elaborate on when one tailed and when two tailed tests are used.

7.7 The following readings give the weights of four bags of sugar in kg. from a shipment: 100, 105, 95, 98. Find estimates of the following:

- (i) The population means weight of all the bags.
- (ii) The standard deviation of the weights of all the bags.

7.8 10 randomly selected people were asked whether they had seen the latest advertisement of an air conditioner. The following responses were recorded “yes”, “yes”, “no”, “yes”, “no”, “no”, “yes”, “no”, “yes”, “yes”. Obtain a point estimate of the proportion of people who have seen the latest air conditioner advertisement.

7.9 There are two soft drink bottling machines, A and B. We are interested in the fill and take two random samples of 50 bottles each from each machine and calculate the following:

Machine	A	B
\bar{x}	19.5	19.3
σ	1.4	1.8

- (a) Estimate the difference in average fill between the two machines with a 95% confidence interval.
- (b) How does this compare with previous results?

7.10 There are two vending machines dispensing soft drinks. Measurements of fill weights for 10 randomly chosen drinks from each machine yield the following statistics on fill weight:

Machine	A	B
\bar{x}	7.8 oz	8.2 oz
s	0.4 oz	0.3 oz

Assuming that fill weight follows a normal distribution, estimate the difference in average fill weight between the two machines with a 95% confidence level.

What does the result imply?

- 7.11 In an electoral poll, a study was made by using a sample of 1000 registered voters to find out proportion of voters supporting Congress. In this sample 625 votes said that they would vote for the Congress and the remaining 375 would vote for the BJP. Compute the point estimate of the proportion of all registered voters who would vote for the Congress.
- 7.12 The mean lifetime of a sample of 100 light tubes produced by a company is found to be 1580 hours with standard deviation of 90 hours. Test the hypothesis that the mean lifetime of the tubes produced by the company is 1600 hours.
- 7.13 Suppose a hospital uses large quantities of packaged doses of a particular drug. The individual dose of this drug is 100 cubic cc. The hospital has purchased this drug from the same manufacturer for a number of years and knows that the population SD is 2 cc. The hospital inspects 50 doses of this drug at random from a very large shipment and finds the mean of these doses to be 99.75cc. The hospital sets a 0.10 significance level and asks us whether the doses in this shipment are too small, how can we find the answer?

- 7.14 Is the temperature required to damage a computer on the average less than 110 degrees? Because of the price of testing, twenty computers were tested to see what minimum temperature would damage the computer. It was observed from the sample that the damaging temperature averaged 109 degrees with a standard deviation of 3 degrees. (use 5 % level of significance)
- 7.15 The prices of shares (in Rs) of a company on the different days in a month were found to be 66, 65, 69, 70, 69, 71, 70, 63, 64, 68. Test whether the mean price of the share in the month is more than 65.
- 7.16 In a sample of 750 people, 27% said they feel that health care is the most important issue facing our state. What proportion of the population feels that health care is the most important issue?
- 7.17 A random sample of size $n=20$ is drawn from a normal distribution with standard deviation 6. The sample mean is calculated to be 45. Using these information find 95% confidence interval for the true population mean.
- 7.18 The average monthly rent for a one-bedroom apartment in a city is known to be Rs.2000 with a standard deviation of Rs.500. A random sample of 400 households of one-bedroom apartments showed an average monthly rent of Rs.2500. At $\alpha = + 0.01$, can we conclude that the sample reasonably represents the population?
- 7.19 A manufacturer of certain brand of bulbs claims that the average life of the bulb is 400 hours with a standard deviation of 5 hours. To test the manufacturer's claims, a random sample of 100 bulbs was tested and it showed an average life of 380 hours. What can you conclude about the manufacturer's claim at a level of significance of $\alpha = + 0.05$?
- 7.20 A luggage manufacturing company claims that 80% of executives of a certain company carried briefcases produced by them. A random sample of 900 executives showed that 675 of them carried these briefcases. At 0.05 level of significance, verify the company's claims?
- 7.21 A producer of soft drinks claims that 35 percent of all drinkers prefer its product. A competitor wants to test this claim. A random sample of 200 soft drinkers was taken at random and it was found that 62 of them preferred the producer's brand. At $\alpha = + 0.01$, what can the competitor conclude about the producer's claim?
- 7.22 A customer is trying to determine which of the two fast food burger restaurants, A or B gives faster service. The average waiting time at restaurant A over a period of 60 randomly selected days between the time of 4p.m. and 5p.m. is 6 minutes with a standard deviation of 1.9 minutes. The average waiting time at the restaurant B for a randomly selected 40 day period is found to be 8 minutes between the same time period of 4 p.m. and 5p.m., with a standard deviation of 3.2 minutes. Can the customer conclude that one of the restaurants gives faster service than the other, at 0.01 level of significance?
- 7.23 A marketing research firm was asked by a major television network to conduct a survey to study how much time the heads of households in different parts of the country spend on leisure activities on Saturday. A random sample of 60 heads of households each from Bangalore and Delhi was selected, and the time spent on leisure activities in terms of watching TV or reading books or other non – professional activities was noted. The results are shown as follows:

Bangalore	Delhi
$\bar{x}_1 = 8$ hours	$\bar{x}_2 = 9$ hours
$s_1 = 2.0$ hours	$s_2 = 2.5$ hours

- (a) At 95 percent confidence level, is there evidence that the average amount of time spent on leisure by heads of households in Bangalore is greater than 7 hours?
- (b) At 99 percent confidence level, is there evidence of a difference in the average time spent between heads of households in Bangalore and Delhi?
- 7.24 Random samples of 3000 people in Delhi and 4000 Mumbai were asked if they thought there was too much violence shown on TV these days. 1500 people in Delhi and 1800 people in Mumbai replied in the affirmative. Can we conclude at 99% confidence level that the two proportions are significantly different?
- 7.25 A marketing major student at Delhi University is interested in a survey to find out if women are more likely to use credit cards while shopping in a department store. She took random samples of shopping records of 100 men and 100 women at the local shopping malls. 75 men and 80 women had made their purchases by using credit cards on a given day. At 0.05 level of significance, what conclusion can be drawn?
- 7.26 A study was recently conducted to find out if there was a significant difference between the proportion of married men and married women who had achieved the level of CEO. In a random sample of 40 male CEO's and 20 female CEO's of 60 fortune 500 companies, 90 percent of male CEO's and 55 percent of female CEO's were married. At 90 percent confidence level, can we conclude that fewer married women would advance or remain at CEO level than men?
- 7.27 A consignment of cricket bats contained 10,000 bats. Out of this lot, a sample of 100 cricket bats was drawn and out of these 5 were found to be defective. Construct a 90% confidence interval for the true proportion of defective bats in the consignment.
- 7.28 A toothpaste pump machine fills toothpaste tubes to 100gms. The manufacturer is interested in estimating an interval for the true mean of the toothpaste tubes. A sample of 500 toothpaste tubes gave the mean as 101gms. Construct a 95% confidence interval of the true mean amount of toothpaste in the tubes.
- 7.29 A newspaper is conducting a telephonic survey on the response of people to a new legislation passed in parliament. 60% of the people responded in favor of the legislation. Find a 90% confidence interval for the true proportion of people who support the new legislation.
- 7.30 A factory produces 50,000 watches daily. A sample of 500 watches was drawn and 5% was found to be defective. Find a 95% C.I. for the true proportion of defective watches in the factory.
- 7.31 A company has developed a new drug (X) for curing common cold, which they believe, is more effective than the existing drug (Y). The new drug X was administered on 200 patients suffering from cold and drug Y on 250 patients suffering from cold. 70% of the people responded positively to drug X and 65% to drug Y. Construct a 95% confidence interval for the difference in the true proportion of patients who might be expected to respond to the two drugs.
- 7.32 In a random sample of 100 persons taken from village A, 60 are found to be consuming tea. In another sample of 200 persons taken from village B, 100 persons are found to be consuming tea. Do the data reveal significant difference between the two villages so far as the habit of taking tea is concerned? **(MBA, DU, 1999)**
- 7.33 Before an increase in exercise duty on tea, 400 people out of a sample of 500 people were found to be tea drinkers. After an increase in duty, 400 people were tea drinkers in a sample of 600 people. State, whether there is a significant decrease in the consumption of tea. **(MBA, DU, 2002)**

7.34 Intelligence test given to two groups of boys and girls gave the following information:

	Mean Score	S.D.	Number
Girls	75	10	50
Boys	70	12	100

Is the difference in the mean scores of boys and girls statistically significant?

(MBA, S.V. Univ., 1995; MBA, DU, 1997)

7.35 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 is used in the other area. In a random sample of 60 customers who saw advertisement A, 18 tried the product. In a random sample of 100 customers who saw advertisement B, 22 tried the product. Does this indicate that advertisement A is more effective than advertisement B, if a 5% level of significance is used?

(MFC, DU, 1996; MBA, DU, 2000, MBA, IGNOU 2002)

7.36 500 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 at 5% level of significance that production at second factory is better than in first factory?

(MBA, Kurukshetra Univ., 1996; MBA, DU, 2002)

7.37 The mean of two random samples of sizes 9 and 7 are 196.42 and 198.82 respectively. The sum of 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? **(DU, 2004)**



8

Correlation and Regression



Structure

8.1 Correlation Analysis

8.1.1 Graphical Representation of Correlation

8.1.2 Covariance

8.1.3 Correlation Coefficient

8.1.3.1 Karl Pearson's Correlation Coefficient

8.1.3.2 Properties of Correlation Coefficient

8.1.3.3 Standard Error and Probable Error of Correlation Coefficient

8.1.4 Coefficient of Determination

8.1.5 Rank Correlation

8.1.6 Partial Correlation

8.1.7 Multiple Correlation

8.1.8 Testing the Significance of Correlation Coefficient

8.1.9 Testing the Significance of Partial Correlation Coefficient

8.1.10 Testing the Significance of Multiple Correlation Coefficient

8.2 Regression Analysis

8.2.1 Simple Linear Regression

8.2.1.1 Regression Equation and Regression Coefficients

8.2.1.2 Properties of Regression Coefficients

8.2.1.3 Least Square Method and Regression Equation

8.2.1.4 Explained and Unexplained Variation

8.2.1.5 Standard Error of Estimate

8.2.1.6 Testing the Significance of Regression Coefficients

8.2.2 Multiple Regression

8.2.2.1 Multiple Regression with two Independent Variables

8.2.2.2 Regression with Dummy Variable

8.3 Caselet

8.4 Excel Guide

8.5 Exercises

INTRODUCTION

In the earlier chapters we have discussed how to summarize the characteristics of a single variable. The next question is how to summarize a pair of bi-variate variables measured on the same unit. How do we describe their joint behavior? This chapter begins with the study of describing data that contain more than one variable. Correlation and Regression analysis are the two basic methods to arrest this characteristic of bivariate data. In section 8.1 we will see how the correlation coefficient and scatter diagram are used to describe bivariate data. Regression analysis is discussed in section 8.2.

8.1 CORRELATION ANALYSIS

Correlation is a statistical technique, which can show whether, and how strongly pairs of variables are related. For example, height and weight are related - taller people may tend to be heavier than shorter people. The relationship isn't perfect. People of the same height vary in weight, and we can easily think of two people one knows where the shorter one is heavier than the taller one. Nonetheless, the average weight of people 5'5" is less than the average weight of people 5'6", and their average weight is less than that of people 5'7", etc. Correlation can tell just how much of the variation in peoples' weights is related to their heights. Correlation is a bivariate measure of association (strength) of linear relationship between two variables. Thus, correlation analysis is a statistical tool which can be used to describe the degree to which one variable is linearly related to another.

If a linear relationship exists, a correlation co-efficient would indicate how close to a straight line, such a relationship is and this is known as the strength of the association.

Once linear relationship is established by correlation analysis, regression analysis takes it a step further by quantifying and establishing a straight line relationship. This straight line can be further used for prediction purposes i.e. it is possible to predict one variable given that values of the other variable are known. The variable that is predicted is known as the dependent variable or response variable and the variable that is used for making the prediction is known as the independent variable.

There are several different correlation techniques. The present analysis includes the most common kinds like Karl Pearson's or product-moment correlation, Spearman Rank Correlation etc.. The concept of partial correlation and multiple correlation are also briefly discussed here.

Like all statistical techniques, correlation is only appropriate for certain kinds of data. Correlation works for data in which numbers are meaningful, usually quantities of some sort. It cannot be used for purely categorical data, such as gender, brands purchased or favorite color.

8.1.1 Graphical Representation of Correlation

As already mentioned above, correlation indicates the degree of association between two variables. Graphically this relation can be presented by the Scatter Diagram. Scatter diagram is mainly used to represent bivariate data. Although these diagrams cannot prove that one variable causes changes in the other, they do indicate the existence of a relationship, as well as the strength of that relationship.

A scatter diagram is composed of a horizontal axis containing the measured values of one variable and a vertical axis representing the measurements of the other variable.

Steps of drawing the Scatter diagram

1. Collect data on two variables, one independent and the other dependent.

2. Draw a diagram labelling the horizontal and vertical axes. It is common that the “cause” variable or independent variable be labeled on the horizontal (X) axis and the “effect” variable or the dependent variable be labeled on the vertical (Y) axis. The values should increase up the vertical scale and toward the right on the horizontal scale. The scale on both the X and Y-axes should be sufficient to include both the largest and the smallest X and Y values in the table.
3. Plot the data pairs on the diagram by placing a dot at the intersections of the X and Y coordinates for each data pair.

INTERPRETATION OF SCATTER DIAGRAM

The relationship may be positive, negative or may even display no relationship.

- (i) A positive relationship is indicated by an ellipse of points that slopes upward demonstrating that an increase in the cause variable also increases the effect variable.
- (ii) A negative relationship is indicated by a collection of points that slopes downward demonstrating that an increase in the cause variable results in a decrease in the effect variable.
- (iii) A diagram with a cluster of points such that it is difficult or impossible to determine whether the trend is upward sloping or downward sloping indicates that there is no relationship between the two variables.
- (iv) The strength of the relationship can also be examined from the closeness of the clustered points. The more the points are clustered to look like a straight line the stronger the relationship.

The following figure 8.1 indicates the different types of association between two variables X and Y.

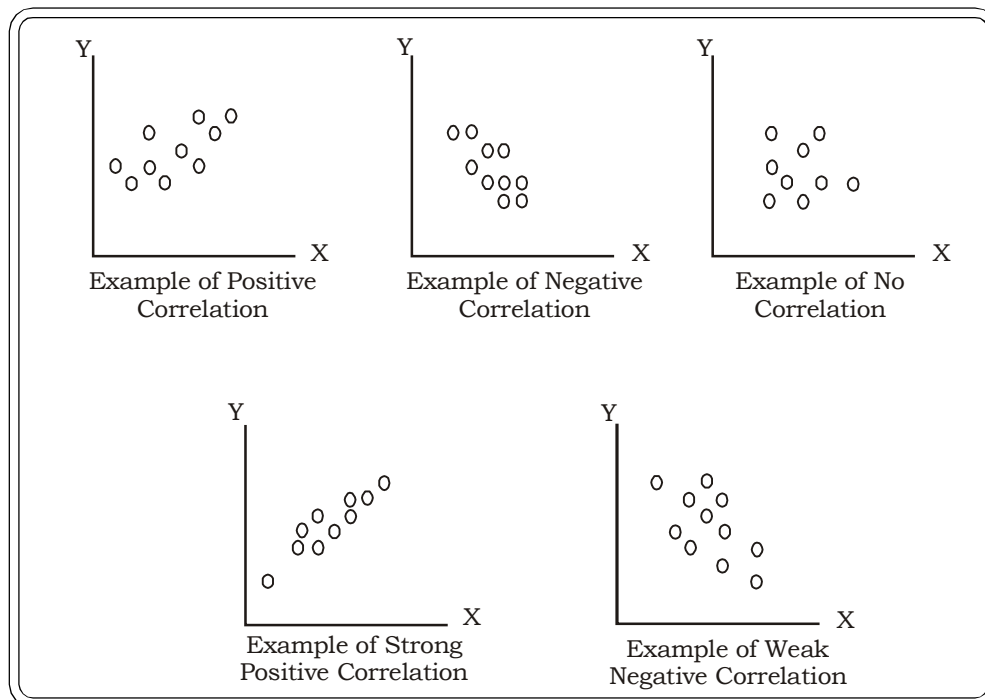


Figure 8.1

Different Kinds of Correlation

8.1.2 Covariance

Covariance is a measure of how much the deviations of two variables match. It can be expressed by the following formula:

For two variables X and Y,

$$\begin{aligned} \text{cov}(X, Y) &= \frac{1}{n} \sum (X - \bar{X})(Y - \bar{Y}) \\ &= \frac{\sum XY}{n} - \left(\frac{\sum X}{n} \right) \left(\frac{\sum Y}{n} \right) \end{aligned}$$

Covariance has the same properties as that of variance. It is independent of change in origin but dependent of change in scale i.e.

$$\begin{aligned} \text{if } U &= \frac{X-A}{I} \text{ and } V = \frac{Y-B}{J} \\ \text{cov}(U, V) &= IJ \text{cov}(X, Y) \end{aligned}$$

8.1.3 Correlation Coefficient

Given a pair of related measures (X and Y) on each of a set of items, the correlation coefficient (r) provides an index of the degree to which the paired measures co-vary in a linear fashion. The correlation between two variables reflects the degree to which the variables are linearly related. The most common measure of correlation is Karl Pearson's Product Moment Correlation (popularly called Pearson's Correlation Coefficient). When measured in a population, the Pearson Product Moment correlation is designated by the Greek letter rho (ρ). When computed in a sample, the letter "r" designates it.

8.1.3.1 Karl Pearson's Correlation Coefficient

Karl Pearson's correlation coefficient is a quantity that gives the amount of linear relationship between two variables. Mathematically, it is the ratio of the covariance of two variables, divided by the product of their standard deviations. Thus for two variables X and Y, Karl Pearson's correlation co-efficient is given by,

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

where cov (X, Y) is the covariance between X & Y, σ_X and σ_Y are the standard deviations of X and Y respectively.

$$\text{cov}(X, Y) = \frac{1}{n} \sum XY - \frac{1}{n} \sum X \sum Y$$

$$\sigma_x^2 = \frac{1}{n} \sum X^2 - \bar{X}^2$$

$$\sigma_y^2 = \frac{1}{n} \sum Y^2 - \bar{Y}^2$$

Other forms of Pearson's formula

Karl Pearson's correlation co-efficient (r) takes on many forms some of which are as follows:

$$(1) \quad r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

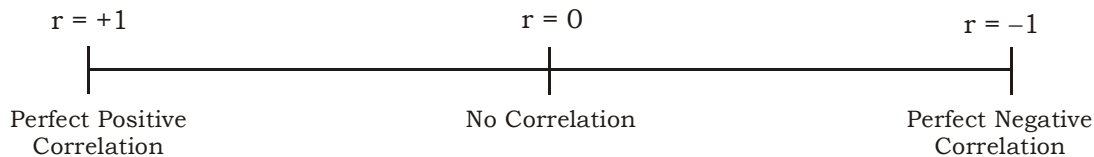
$$r = \frac{\sum (XY - n\bar{X}\bar{Y})}{\sqrt{(\sum X^2 - n\bar{X}^2)(\sum Y^2 - n\bar{Y}^2)}}$$

$$r = \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}} \quad \text{where } x = X - \bar{X} \text{ and } y = Y - \bar{Y}$$

(2) The following formula is also often used to simplify calculations for correlation coefficient.

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$$

Interpretation of Karl Pearson's Correlation Coefficient



- (i) The value of correlation coefficient (or "r"), ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related.
- (ii) If r is positive, it means that as one variable increases the other also increases. If r is negative it means that as one increases, the other decreases (often called an "inverse" correlation).
- (iii) A correlation of +1 means that there is a perfect positive linear relationship between variables. The scatter plot shown below in figure 8.2 depicts such a relationship. It is a positive relationship because high scores on the X-axis are associated with high scores on the Y-axis.

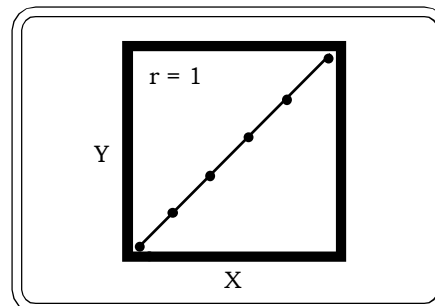


Figure 8.2

$r = 1$: Perfect Positive Correlation

- (iv) A correlation of -1 means below that there is a perfect negative linear relationship between variables. The scatter plot shown depicts a negative relationship. It is a negative relationship because high scores on the X-axis are associated with low scores on the Y-axis.

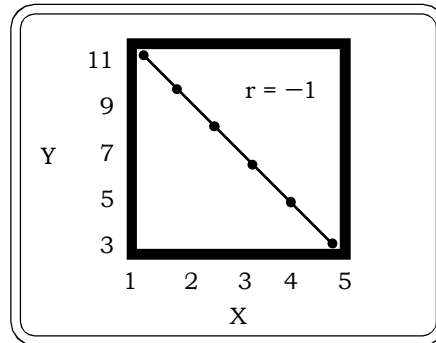


Figure 8.3

$r = -1$: Perfect Negative Correlation

- (v) A correlation of 0 means there is no linear relationship between the two variables. The graph in figure 8.4 shows a Pearson correlation of 0 . Correlations are rarely if ever 0 , 1 , or -1 .

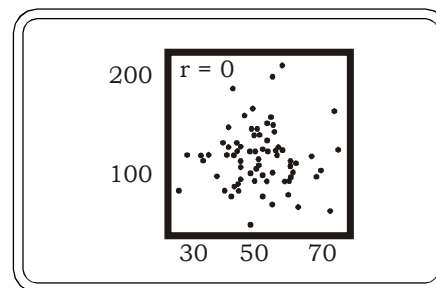


Figure 8.4

$r = 0$: No Correlation

- (vi) While correlation coefficients are normally reported as $r =$ (a value between -1 and $+1$), squaring them makes them easier to interpret. The square of the coefficient (or r square) is equal to the percent of the variation in one variable that is related to the variation in the other. This is called the coefficient of determination and is dealt with in section 8.1.4.

Pearson correlation technique works best with linear relationships: as one variable gets larger, the other gets larger (or smaller) in direct proportion. It does not work well with curvilinear relationships (in which the relationship does not follow a straight line). An example of a curvilinear relationship, consider the variables age and health care. They are related, but the relationship doesn't follow a straight line. Young children and older people both tend to use much more health care than teenagers or young adults. Multiple regression can be used to examine such curvilinear relationships.

8.1.3.2 Properties of Correlation Coefficient

1. Correlation coefficient measures the strength or degree of linear relationship.
2. The value of r lies between $+1$ and -1

$$-1 \leq r \leq 1$$

If

$r = 1 \Rightarrow$ Perfect positive correlation

$r = -1 \Rightarrow$ Perfect negative correlation

$r = 0 \Rightarrow$ No correlation

3. Correlation coefficient is independent of both change in origin and change in scale.

i.e. if $U = \frac{X-A}{I}$ and $V = \frac{Y-B}{J}$

then $r_{xy} = r_{uv}$

4. $r_{xy} = r_{yx} = r$ i.e. relation between x and y is same as the correlation between y and x. Thus correlation coefficient is symmetric.

8.1.3.3 Standard Error and Probable Error of Correlation Coefficient

If r is the correlation co-efficient between two variables X and Y , for a sample of n observations, then the standard error of the correlation coefficient r is given by:

$$\text{S.E. (r)} = \frac{1-r^2}{\sqrt{n}}$$

The Probable Error of the correlation coefficient is given by:

$$\begin{aligned} \text{P.E. (r)} &= 0.6745 \text{ S.E. (r)} \\ &= 0.6745 \frac{1-r^2}{\sqrt{n}} \end{aligned}$$

The Probable Error, in turn gives the limits within which the population correlation coefficient is expected to lie and these limits are:

$$r \pm \text{P.E. (r)}$$

Also,

(i) If $r < \text{P.E. (r)}$, then the correlation is not considered to be significant.

(ii) If $r > 6 \text{ P.E. (r)}$, then correlation between the variables is considered to be significant.

8.1.4 Coefficient of Determination

The coefficient of determination is the square of the Pearsonian correlation coefficient, r^2 as already mentioned. Without going into the detailed mathematics, r^2 may simply be defined as:

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

Explained variation and total variation is described in detail in section 8.2.1.4. It represents the percentage of the variation in the dependent variable explained by the independent variable.

For example, if $r_{xy} = 0.7$, $r^2 = 0.49$. This implies that 49% of the variation in the dependent variable can be attributed to the independent variable. In other words 49% of the variability has been explained and the remaining 51% is unaccounted for. The coefficient of determination is an indicator of how well the model fits the data. In general, r^2 lies between 0 and 1. A r^2 value close to one indicates that all the variability in the dependent variable is well accounted for by the independent variable.

Example: 8.1: The following Table shows 10 years data of advertisement expenditure and sales of a company. Calculate the correlation coefficient between these two variables for this company? Also calculate the coefficient of determination and interpret it.

S.No.	Ad. Expenditure	Sale
1	50	700
2	50	650
3	50	600
4	40	500
5	30	450
6	20	400
7	20	300
8	15	250
9	10	210
10	5	200

Solution:

We have to calculate r i.e. the correlation coefficient, between advertising expenditure (X) and sales (Y)

With the given information r can be calculated by the following formula:

$$r_{xy} = \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}} \text{ where } x = X - \bar{X} \text{ and } y = Y - \bar{Y}$$

$$n = 10$$

S.No.	Ad. Expenditure (X)	Sale (Y)	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	y^2	xy
1	50	700	21	274	441	75076	5754
2	50	650	21	224	441	50176	4704
3	50	600	21	174	441	30276	3654
4	40	500	11	74	121	5476	814
5	30	450	1	24	1	576	24
6	20	400	-9	-26	81	676	234
7	20	300	-9	-126	81	15876	1134
8	15	250	-14	-176	196	30976	2464
9	10	210	-19	-216	361	46656	4104
10	5	200	-24	-226	576	51076	5424
Total	290	4260	0	0	2740	306840	28310

From the above table we can calculate the required terms:

$$\bar{X} = \frac{290}{10} = 29$$

$$\bar{Y} = \frac{4260}{10} = 426$$

$$\sum xy = 28310$$

$$\sum x^2 = 2740$$

$$\sum y^2 = 306840$$

$$\text{Now } r = \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}} = \frac{28310}{\sqrt{2740 \times 306840}} = 0.976$$

Here the value of r indicates high positive correlation between advertisement expenditure and sales of the company.

Coefficient of Determination

$$r^2 = 0.95$$

Thus 95% of sales variation is explained by advertising expenditure.

Example: 8.2: Calculate the correlation coefficient from the following data:

Export of raw cotton (Rs. Crores)	42	44	58	55	89	98	66
Export of manufactured goods (Rs. Crores)	56	49	53	58	65	65	58

Solution:

Export(x)	Import(Y)	u = x-55	v = y-58	u ²	v ²	uv
42	56	-13	-2	169	4	26
44	49	-11	-9	121	81	99
58	53	3	-5	9	25	-15
55	58	0	0	0	0	0
89	65	34	7	1156	49	238
98	76	43	18	1849	324	774
66	58	11	0	121	0	0
		$\sum u = 67$	$\sum v = 9$	$\sum u^2 = 3425$	$\sum v^2 = 483$	$\sum uv = 1122$

$$n = 7$$

Since correlation is independent of change of origin we define $u = x - 55$ and $v = y - 58$

Thus,

$$r_{xy} = r_{uv} = \frac{\text{cov}(uv)}{\sigma_u \sigma_v}$$

$$\begin{aligned} \text{cov}(uv) &= \frac{\sum uv}{n} - \frac{\sum u}{n} \frac{\sum v}{n} \\ &= \frac{1122}{7} - \frac{67}{7} \frac{9}{7} \\ &= \frac{7854 - 603}{49} = \frac{7251}{49} = 147.98 \end{aligned}$$

$$\begin{aligned} \sigma_u &= \sqrt{\frac{\sum u^2}{n} - \left(\frac{\sum u}{n}\right)^2} \\ &= \sqrt{\frac{3425}{7} - \left(\frac{67}{7}\right)^2} = \sqrt{\frac{23975 - 4489}{49}} = \sqrt{\frac{19486}{49}} = 19.94 \end{aligned}$$

$$\sigma_v = \sqrt{\frac{\sum v^2}{n} - \left(\frac{\sum v}{n}\right)^2}$$

$$\sqrt{\frac{483}{7} - \left(\frac{9}{7}\right)^2} = \sqrt{\frac{3381 - 81}{49}} = \sqrt{\frac{3300}{49}} = 8.20$$

$$\text{Thus } r_{xy} = \frac{\text{cov}(uv)}{\sigma_u \sigma_v} = \frac{147.98}{19.94 \times 8.20} = \frac{147.98}{163.51} = 0.91$$

Example: 8.3: Mr. X is the team leader of the MNC Company of Kolkata. His annual income and investment for last six years is given in the following table:

Year	2000	2001	2002	2003	2004	2005
Premium Income (Rs. Lakhs)	4.4	4.8	5.2	5.7	5.9	6.0
Investment (Rs. Lakhs)	1.4	1.8	2.2	2.8	3.0	3.4

Calculate the correlation between Mr. X's income and investment for the last six years.

Solution:

Here we will use Karl Pearson's correlation coefficient as follows:

$$r_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

Income (x)	Investment (y)	(x - \bar{x})	(x - \bar{x}) ²	(y - \bar{y})	(y - \bar{y}) ²	(x - \bar{x}) (y - \bar{y})
4.4	1.4	-0.93	0.86	-1.03	1.06	0.96
4.8	1.8	-0.53	0.28	-0.63	0.40	0.33
5.2	2.2	-0.13	0.02	-0.23	0.05	0.03
5.7	2.8	0.37	0.14	0.37	0.14	0.14
5.9	3.0	0.57	0.32	0.57	0.32	0.32
6.0	3.4	0.67	0.45	0.97	0.94	0.65
			2.07		2.91	2.43

$$\bar{x} = \frac{\sum x}{n} = 5.33$$

$$\bar{y} = \frac{\sum y}{n} = 2.43$$

$$r = \frac{2.43}{\sqrt{2.07 \times 2.91}} = \frac{2.43}{2.45} = 0.99$$

The correlation coefficient is 0.99

Thus, there is high positive correlation between the income and investment of Mr. X.

Example: 8.4: While calculating the coefficient of correlation between U and V, the following results were obtained.

$$n = 25, \sum U = 125, \sum V = 100, \sum U^2 = 650, \sum V^2 = 460, \sum UV = 508.$$

Later on during checking it was observed that two pairs of observation were mistakenly taken as (6, 14) and (8, 6) while the correct data were (8, 12) and (6, 8) respectively. Determine the correct correlation coefficient.

Solution:

Corrected values can be calculated as follows:

$$\sum U = 125 - (6 + 8) + (8 + 6) = 125$$

$$\sum V = 100 - (14 + 6) + (12 + 8) = 100$$

$$\sum U^2 = 650 - (6^2 + 8^2) + (8^2 + 6^2) = 650$$

$$\sum V^2 = 460 - (14^2 + 6^2) + (12^2 + 8^2) = 208$$

$$\sum UV = 508 - (6 \times 14) - (8 \times 6) + (8 \times 12) + (6 \times 8) = 520$$

With these given information r can be calculated by the following formula:

$$r = \frac{\text{cov}(U, V)}{\sigma_U \sigma_V}$$

$$\text{cov}(u, v) = \frac{\sum UV}{n} - \left(\frac{\sum U}{n} \right) \left(\frac{\sum V}{n} \right)$$

$$= \frac{520}{25} - \left(\frac{125}{25} \right) \left(\frac{100}{25} \right) = 20.8 - 20 = 0.8$$

$$\sigma_U = \sqrt{\frac{\sum U^2}{n} - \left(\frac{\sum U}{n} \right)^2} = 1$$

$$\sigma_V = \sqrt{\frac{\sum V^2}{n} - \left(\frac{\sum V}{n} \right)^2} = 1.2$$

$$\text{Therefore } r(\text{corrected}) = \frac{\text{cov}(U, V)}{\sigma_U \sigma_V} = \frac{.8}{1 \times 1.2} = \frac{2}{3} = 0.67$$

Example 8.5: The following data relate to the prices and supplies of a commodity during a period of eight years:

Price	10	12	18	16	15	19	18	17
Supply	30	35	45	44	42	48	47	46

Calculate the coefficient of correlation between the two series.

Solution:

Price (x)	Supply (y)	$x' = x - 16$	$y' = y - 44$	$(x' - \bar{x}')^2$	$(y' - \bar{y}')^2$	$(x' - \bar{x}')(y' - \bar{y}')$
10	30	-6	-4	31.58	146.89	68.11
12	35	-4	-9	13.10	50.69	25.77
18	45	2	1	5.66	8.29	6.85
16	44	0	0	0.14	3.53	.71
15	42	-1	-2	.38	0.01	.07
19	48	3	4	11.42	34.57	19.87
18	47	2	3	5.66	23.81	11.61
17	46	1	2	1.90	15.05	5.35
			69.88	282.88	138.38	

Since correlation coefficient is independent of change in origin.

$$r_{xy} = r_{x'y'}$$

However, the shift at origin will minimize the labour of computation.

$$r_{xy} = \frac{\sum (x' - \bar{x}')(y' - \bar{y}')}{\sqrt{\sum (x' - \bar{x}')^2 \sum (y' - \bar{y}')^2}} = \frac{138.38}{\sqrt{69.88 \times 282.88}} = 0.98$$

Thus, the calculated correlation coefficient is 0.98 which indicates high positive correlation between the price and quality supplied of the given commodity.

8.1.5 Rank Correlation

Besides Karl Pearson's correlation, another important measure of correlation is the rank correlation coefficient proposed by Charles Edward Spearman in 1904. It is a nonparametric or distribution-free rank statistic, which can be used as a measure of the strength of the association between two variables. The Spearman rank correlation coefficient can be used when the distribution of the data is such that it is not possible to quantify it but only rank it in a certain order on the basis of a certain attribute.

The Spearman rank correlation coefficient is defined by the following formula:

$$R = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

where

R - Rank correlation co-efficient

R_{1i} - Rank of the i^{th} X observation

R_{2i} - Rank of the i^{th} Y observation

$$\text{Then } d_i = R_{1i} - R_{2i}$$

= Difference of ranks of the i^{th} X observation and the i^{th} Y observations

$n \rightarrow$ number of pairs of observations

Remarks

- (i) When ranks of the two sets of observation are given, then the formula may be applied directly.
- (ii) When observations are not ranked, then the ranks have to be first assigned by giving the highest observation in the two data set the rank 1 and then ranking the remaining observations accordingly.
- (iii) The case of Tied/ Repeated Ranks.

Often, the case may arise when two or more individuals get the same rank with respect to either of the two characteristics being studied. In such a case, a common rank is assigned to the observations that are repeated. This common rank is the average of the ranks which these observations would have assumed had they been different from one another.

For example, if two observations were ranked equal at the fourth place, then both these two observations would be assigned the rank.

$$\frac{4 + 5}{2} = 4.5$$

and the next observation would be ranked 6 and so on.

Also, equivalently an adjustment or correction factor is applicable in the formula for each observation that is repeated.

The adjusted formula for rank correlation coefficient is:

$$R = 1 - \frac{6 \left\{ \sum_{i=1}^n d_i^2 + \frac{1}{12} m_1(m_1^2 - 1) + \frac{1}{12} m_2(m_2^2 - 1) + \dots \right\}}{n(n^2 - 1)}$$

where m_i = The number of times the i^{th} repeated item is repeated

$i = 1, 2, \dots$

Example 8.6: A group of 5 Army officers have participated in the competition of both SWIMMING and RUNNING. The following table depicts the ranks, which is in accordance with the achievements in both the tests. Find out the relationship between the performances of the officers in the two events.

Officer	Running	Swimming
Mr. Abhijit	5	3
Mr. Atul	2	1
Mr. Manish	4	5
Mr. Dipak	1	2
Mr. Gagan	3	4

Solution:

To answer the above question we have to consider the Rank Correlation Coefficient of Spearman as the data are in ordinal form. The formula, as already discussed, is

$$R = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where, $d = R_1 - R_2$,

where, $R_1 \rightarrow$ Rank in Running

where, $R_2 \rightarrow$ Rank in Swimming

This is the case when ranks have already been assigned.

All that is required for the calculation of the Spearman coefficient are the values of n and $\sum d^2$. In the table below, differences between each pair of ranks i.e. d_i ($d_i = R_1 - R_2$) and the sum of d^2 have been calculated.

Officer	Running (R_1)	Swimming (R_2)	d_i = $R_1 - R_2$	d_i^2
Mr. Abhijit	5	3	2	4
Mr. Atul	2	1	1	1
Mr. Manish	4	5	-1	1
Mr. Dipak	1	2	-1	1
Mr. Gagan	3	4	-1	1
				$\sum d^2 = 8$

From the above table we have,

$$\sum d_i^2 = 4 + 1 + 1 + 1 + 1 = 8$$

$$n = 5; n^2 = 25$$

We have,

$$\begin{aligned} R &= 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \\ &= 1 - \frac{6 \times 8}{5(25 - 1)} \\ &= 1 - \frac{2}{5} = \frac{3}{5} = 0.6 \end{aligned}$$

The correlation is 0.6. Thus the relationship is positive but not very strong.

Example 8.7: Suppose, two tea testers are asked to rank 8 types of tea from best to worst (rank # 1 = best, rank # 8 = worst) with respect to their fundamental importance or whatever else it might be that strikes the testers' fancy. Their ranking is shown in the table below. Find out whether the testers are similar in nature or not in the present case.

Tea Variety	Rank by Tester 1	Rank by Tester 2
1	1	2
2	2	1
3	3	5
4	4	3
5	5	4
6	6	7
7	7	8
8	8	6

Solution:

To answer the above question, we consider the Spearman's Rank Correlation Coefficient. The formula is

$$R = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

where $d = R_1 - R_2$

where R_1 are the ranks assigned by Tester 1

and R_2 are the ranks assigned by Tester 2

Tea Variety	Rank by Tester 1	Rank by Tester 2	d	d ²
1	1	2	-1	1
2	2	1	1	1
3	3	5	-2	4
4	4	3	1	1
5	5	4	1	1
6	6	7	-1	1
7	7	8	-1	1
8	8	6	2	4
n = 8				$\sum d^2 = 14$

From the above table we have,

$$\sum d^2 = 14$$

$$\text{Thus } R = 1 - \frac{6 \times 14}{8(8^2 - 1)} = 1 - \frac{84}{8 \times 63} = 0.83$$

The calculated R is 0.83. So it can be concluded that there is a substantial degree of similarity between the rankings of the two experts.

Example: 8.8: A large manufacturing firm wants to determine whether a relationship exists between the number of works-hours an employee misses per year and the employee's annual wages (in thousands of rupees). A sample of 15 employees produced the data shown in the following table.

Employee	Hours	Wages
1	49	15.8
2	36	17.5
3	127	11.3
4	91	13.2
5	72	13.0
6	34	14.5
7	155	11.8
8	11	20.2

9	191	10.8
10	6	18.8
11	63	13.8
12	79	12.7
13	43	15.1
14	57	24.2
15	82	13.9

Calculate Spearman's rank correlation coefficient as a measure of the strength of the relationship between work-hours missed and annual wages.

Solution:

- (a) First, we rank the values of work-hours missed and rank the values of the annual salaries. Let these rankings be R_1 and R_2 , respectively, and these are shown in the following table.

Employee	Hours	Rank (R_1)	Wages	Rank (R_2)	d_i	d_i^2
1	49	6	15.8	11	-5	25
2	36	4	17.5	12	-8	64
3	127	13	11.3	2	11	121
4	91	12	13.2	6	6	36
5	72	9	13.0	5	4	16
6	34	3	14.5	9	-6	36
7	155	14	11.8	3	11	121
8	11	2	20.2	14	-12	144
9	191	15	10.8	1	14	196
10	6	1	18.8	13	-12	144
11	63	8	13.8	7	1	1
12	79	10	12.7	4	6	36
13	43	5	15.1	10	-5	25
14	57	7	24.2	15	-8	64
15	82	11	13.9	8	-3	9
						$\sum d_i^2 = 1038$

Since there are no ties, we calculate R_s by the formula

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6(1038)}{15(224)} = -0.854$$

This large negative value of R implies that a negative correlation exists between work-hours missed and annual wages in the sample of 15 employees.

Example 8.9: The scores of 10 students on the mid term examination and the final examination are as follows:

Student	Mid-Term Score (x)	Final Exam Score (y)
Neha	82	94
Chani	81	92
Aditi	80	85
Sumit	68	75
Aditya	70	73
Mohit	92	95
Reha	76	69
Rahul	80	86
Sakshi	86	90
Charu	62	69

Compute the rank correlation coefficient.

Solution:

It may be observed that there are repeat values in the data and hence the rankings must be given accordingly. The ranks are shown in the following table.

Mid Term Score (x)	Rank R_1	Final Exam Score (y)	Rank R_2	$d^2 = (R_1 - R_2)^2$
82	3	94	2	1
81	4	92	3	1
80	5.5	85	6	0.25
68	9	75	7	4
70	8	73	8	0
92	1	95	1	0
76	7	69	9.5	6.25
80	5.5	86	5	0.25
86	2	90	4	4
62	10	69	9.5	0.25
				$\Sigma d^2 = 17$

Also, since 80 is repeated twice, $m_1 = 2$ and in the second series 69 is repeated twice. Thus $m_2 = 2$ and $n = 10$.

The rank correlation coefficient adjusted for repeat observations is:

$$R = 1 - \frac{6 \left\{ \Sigma d_i^2 + \frac{1}{12} m_1 (m_1^2 - 1) + \frac{1}{12} m_2 (m_2^2 - 1) \right\}}{n(n^2 - 1)}$$

$$= 1 - \frac{6 \left\{ 17 + \frac{1}{12} 2(2^2 - 1) + \frac{1}{12} \times 2(2^2 - 1) \right\}}{10(10^2 - 1)}$$

$$R = 1 - \frac{6\{17 + 1.92 + 1.92\}}{990}$$

$$= 1 - 0.13$$

$$= 0.987$$

Thus, spearman's Rank correlation coefficient = 0.987

8.1.6 Partial Correlation

Partial correlation is useful when we want to study the relationship between two variables while removing the linear effect of certain other variables. Partial correlation is the correlation of two variables while controlling for a third or more other variables. The technique is commonly used in "causal" modeling of small models (3 – 5 variables). For instance, $r_{12.3}$ is the correlation between

variables 1 and 2, after eliminating the effect of variable 3. The researcher compares the controlled correlation (i.e., $r_{12.3}$) with the original correlation (i.e., r_{12}) and, if there is no difference, the inference is that the control variables have no effect. If the partial correlation approaches 0, the inference is that the original correlation is spurious — there is no direct causal link between the two original variables indicates correlation between variable 1 and 2 keeping variable 3 constant.

Mathematically partial correlation between of variables X_1 and X_2 after eliminating the effect of X_3 , denoted by $r_{12.3}$, can be calculated as follows:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

where

$r_{12} \Rightarrow$ Karl Pearson's correlation coefficient between variable X_1 and variable X_2 .

$r_{13} \Rightarrow$ Karl Pearson's correlation coefficient between variable X_1 and variable X_3 .

$r_{23} \Rightarrow$ Karl Pearson's correlation coefficient between variable X_2 and variable X_3 .

Similarly,

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}}$$

$$\text{and } r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{13}^2}}$$

Remarks

(i) It may be noted that r_{12} , r_{23} and r_{13} are called total correlation coefficients and $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$ are called first order partial correlation co-efficient.

The order of the correlation coefficient is determined by the number of secondary subscripts i.e. the number of subscripts after the dot.

In case of four variables X_1 , X_2 , X_3 and X_4 , for example,

$r_{12.34}$ is a second order partial correlation coefficient between X_1 and X_2 after eliminating the linear effect of X_3 and X_4 .

(ii) Partial correlation coefficient lies between ± 1 .

Example 8.10: Suppose the computer has found for a given set of values.

$$r_{12} = 0.96, r_{13} = 0.36 \text{ and } r_{23} = 0.78.$$

Examine whether the computations may be said to be free from error.

Solution:

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

$$\begin{aligned}
 &= \frac{0.96 - (0.36)(0.78)}{\sqrt{1 - (0.36)^2} \sqrt{1 - (0.78)^2}} \\
 &= \frac{0.96 - 0.2808}{(0.93)(0.62)} \\
 &= \frac{0.6792}{0.5766} \\
 &= 1.17
 \end{aligned}$$

which is greater than 1. Since $r_{12,3}$ exceeds the upper limit, it may be concluded that the computations are not error free.

Example: 8.11: In one study, the correlation between a child's school achievement and the number of hours the child watches TV was -0.33. The correlation between school achievement and teacher ratings of the child's aggressiveness was -0.48. Ratings of aggressiveness and number of hours watching TV correlated 0.55. Find the influence of aggressiveness in the relation of the rest two factors.

Solution:

We can find the correlation between TV watching and school achievement, with ratings of aggressiveness controlled to answer the given question.

The given information can be represented in matrix form as follows:

	School achievement (1)	Hours the child watches TV (2)	Child's aggressiveness (3)
School achievement (1)	1	- 0.33	- 0.48
Hours the child watches TV (2)	- 0.33	1	0.55
Child's aggressiveness (3)	- 0.48	0.55	1

The above matrix is called a correlation matrix.

Let

Variable 1 - School achievement

Variable 2 - Hours the child watches TV.

Variable 3 - Child's aggressiveness.

$$r_{12} = -0.33$$

$$r_{13} = -0.48$$

$$r_{23} = r_{32} = 0.55$$

We have to find out

$$r_{12.3} = ?$$

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} = \frac{(-.33) - (-.48)(.55)}{\sqrt{1-(-.48)^2}\sqrt{1-(.55)^2}} = \frac{-0.66}{.733} = -0.09$$

The partial correlation is -0.09. Thus it can be concluded that, aggressiveness seems to mediate most of the relationship between TV watching and school achievement and when it is partialled out, the correlation between school achievement and hours the child watches TV became less.

Example 8.12: Given that:

X_1 – marks of a group of students in Statistics

X_2 – marks of a group of students in Economics

X_3 – marks of a group of students in English

Given $r_{12} = 0.80$, $r_{23} = 0.60$ and $r_{13} = 0.67$.

Calculate the partial correlation coefficients $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$.

Solution:

$$\begin{aligned} r_{12.3} &= \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} \\ &= \frac{0.80 - (-0.60)(0.67)}{\sqrt{0.55}\sqrt{0.64}} \\ &= \frac{0.80 - 0.402}{(0.74)(0.8)} \\ &= \frac{0.398}{0.592} = 0.672 \end{aligned}$$

$$\begin{aligned} r_{13.2} &= \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}} \\ &= \frac{0.67 - (-0.80)(0.60)}{\sqrt{0.64}\sqrt{0.36}} \\ &= \frac{0.19}{(0.6)(0.8)} \\ &= 0.395 \end{aligned}$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{12}^2}}$$

$$\begin{aligned}
 &= \frac{0.60 - (-0.80)(0.67)}{\sqrt{0.74}\sqrt{0.6}} \\
 &= \frac{0.064}{0.444} \\
 &= 0.144
 \end{aligned}$$

8.1.7 Multiple Correlation

Multiple correlation is an extension of simple correlation. Three or more variables are involved in multiple correlations. If there are three variables X_1 , X_2 and X_3 , the multiple correlation denoted by $R_{1.23}$, is defined as the correlation between X_1 and the joint effect of X_2 and X_3 on X_1 . Normally multiple correlation coefficient for three variables X_i , X_j and X_k is denoted by $R_{i.jk}$. The first subscript denotes the dependent variable and the remaining two independent variables. It is computed by using the formula:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

$R_{1.23}$ indicates correlation between X_1 in one hand and X_2 , X_3 on the other.

$$\text{Similarly } R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}}$$

is the correlation between X_2 and the joint effect of X_1 and X_3 .

$$\text{and } R_{3.12} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}}$$

is the correlation between X_3 and the joint effect of X_1 and X_2 .

Remarks

(i) Generalization

In case of n variables $X_1, X_2, X_3, \dots, X_n$, the multiple correlation coefficient between X_1 and joint effect of X_2, \dots, X_n is denoted by $R_{1.23 \dots n}$

(ii) The range of the multiple correlation coefficient is from 0 to 1.

(iii) If $R_{1.23} = 0$, then all the partial and total correlations are zero i.e. X_1 is completely uncorrelated with all the other variables.

(iv) If $R_{1.23} = 1$, then the association is perfect.

(v) $R_{1.23}$ is not less than any total correlation coefficient i.e. $R_{1.23} \geq r_{12}, r_{13}, r_{23}$.

(vi) The relationship between total, multiple and partial correlations may be expressed as follows

$$1 - R_{1.23}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)$$

Example 8.13: The simple correlations between weight height (r_{12}), weight – age (r_{13}) and height-age is (r_{23}) are given as:

$$r_{12} = 0.98$$

$$r_{13} = 0.44$$

$$r_{23} = 0.54$$

Calculate multiple correlation coefficient taking weight as dependent variable and height and age as independent.

Solution:

Here X_1 : Weight

X_2 : Height

X_3 : Age

Here, we have to calculate $R_{1.23}$ which is defined by the following formula:

$$\begin{aligned} R_{1.23} &= \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \\ &= \sqrt{\frac{(.98)^2 + (.44)^2 - 2(.98)(.44)(.54)}{1 - (.54)^2}} \\ &= \sqrt{\frac{0.69}{0.71}} = 0.99 \end{aligned}$$

The required multiple correlation between weight on one hand and height and age on the other is 0.99.

Example 8.14: For the data given in example 8.12, calculate $R_{1.23}$.

Solution:

$$\begin{aligned} R_{1.23} &= \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{23}^2}} \\ &= \sqrt{\frac{(0.80)^2 + (0.67)^2 - 2(0.80)(0.67)(0.60)}{1 - (0.60)^2}} \\ &= \frac{0.64 + 0.45 - 0.6432}{0.64} \\ &= \frac{0.4468}{0.64} = 0.698 \end{aligned}$$

8.1.8 Testing the Significance of Correlation Coefficient

As we have already discussed in the earlier sections the correlation coefficient can be used to determine whether there is any evidence of significant association between the variables. After calculating the correlation coefficient (r), it is necessary to test its statistical significance.

If the population correlation coefficient of populations is ρ , then the following hypothesis can be formed:

Null Hypothesis:

$H_0: \rho = 0$ i.e. population correlation coefficient is not significant.

Alternative Hypothesis:

$H_1: \rho \neq 0$ i.e. correlation is significant.

The test statistic is as follows:

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2} \sim t(n - 2)$$

The test statistic follows a t distribution with $n-2$ degrees of freedom.

If the value of t is significant i.e. on comparing with the tabulated t at a appropriate level of confidence, if calculated $t >$ tabulated t , then H_0 is rejected. In this case the sample correlation coefficient is considered significant of correlation in the population.

If the value of t is not significant i.e.

calculated $t <$ tabulated t ,

then H_0 may be accepted and in this case sample correlation is not considered indicative of correlation in the population.

8.1.9 Testing the Significance of Partial Correlation Coefficient

We give the test for three variables here (it is possible to generalize this test)

Null Hypothesis

$H_0: \rho_{12.3} = 0$ i.e. population partial correlation co-efficient is zero.

Alternative Hypothesis

$H_1: \rho_{12.3} \neq 0$ i.e. population partial correlation co-efficient is not zero.

Test statistic:

$$t = \frac{r_{12.3} \sqrt{n - k - 2}}{\sqrt{1 - r_{12.3}^2}} \sim t_{n-k-2}$$

where n = number of observations

$r_{12.3}$ = partial correlation coefficient between X_1 and X_2 on eliminating the linear effect of X_3 .

k = order of the partial correlation coefficient as determined by the number of secondary subscripts.

Test criteria

Reject H_0 at $\alpha\%$ level of significance if

$$|t| > t_{n-k-2} \left(\frac{\alpha}{2} \right)$$

else H_0 may be accepted.

8.1.10 Testing the Significance of Multiple Correlation Coefficient

In this section, we describe the test for three variables X_1 , X_2 and X_3 . (The test may also be generalized to more than 3 variables)

Null Hypothesis

$H_0 : \rho_{1,23} = 0$ i.e. population multiple correlation coefficient is zero.

Alternative Hypothesis

$H_1 : \rho_{1,23} \neq 0$ i.e. population multiple correlation coefficient is not zero.

The test statistic

$$F = \frac{R_{1,23}^2}{1 - R_{1,23}^2} \frac{n - k - 1}{k} \sim F(k, n - k - 1)$$

where n = number of observations.

$R_{1,23}$ = multiple correlation between X_1 and joint effect of X_2 and X_3 on X_1 .

k = number of secondary subscripts.

Decision Rule

Reject H_0 at $\alpha\%$ level of significance, if calculated $F >$ tabulated $F(k, n - k - 1, \frac{\alpha}{2})$, else accept H_0 .

Example 8.15: The following data gives the ages of husbands and wives (10 pairs) in years at the time of their marriage:

Husband (x):	23	27	28	29	30	31	33	35	36	39
Wife (y):	18	22	23	24	25	26	28	29	30	32

Compute the correlation coefficient and test its significance.

Solution:

We may use the following formula for calculating the correlation coefficient:

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2} \sqrt{n\sum Y^2 - (\sum Y)^2}}$$

The following quantities are computed from the given data.

$$\begin{aligned} \sum X &= 311, & \sum Y &= 257 \\ \sum X^2 &= 9875, & \sum Y^2 &= 6763, & \sum XY &= 8171 \end{aligned}$$

Thus,

$$\begin{aligned} r &= \frac{10 \times 8171 - (311) \times (257)}{\sqrt{10 \times 9875 - 96721} \sqrt{10 \times 6763 - 66049}} \\ &= 0.9955 \end{aligned}$$

To test the significance of this correlation, we first set up the null and the alternative hypothesis.

$H_0 : \rho = 0$ i.e. population correlation coefficient is zero.

$H_1 : \rho \neq 0$ i.e. population correlation coefficient is not zero.

The test statistic:

$$\begin{aligned} t &= \frac{r}{\sqrt{1-r^2}} \sqrt{n-2} \\ &= \frac{0.9955}{\sqrt{1-0.9955^2}} \sqrt{10-2} \\ &= 29.7371 \end{aligned}$$

$$\text{Tabulated } t \left(\frac{0.05}{2}, 8 \right) = 2.306$$

Decision Rule:

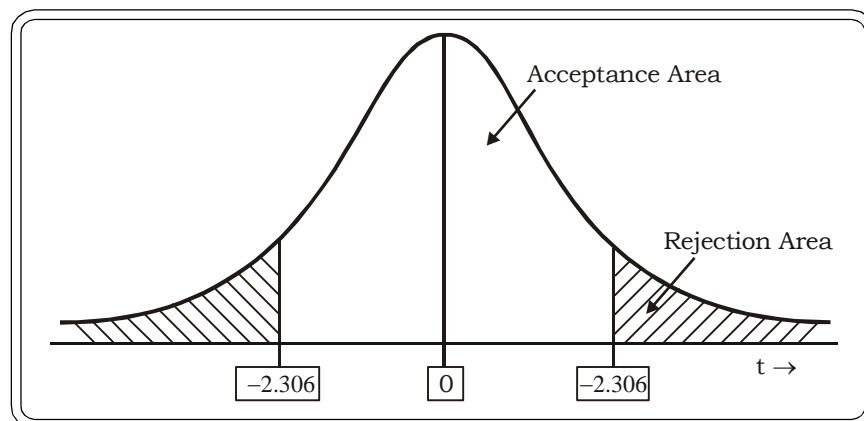
Since calculated $t >$ tabulated t , we may reject the null hypothesis and conclude that $\rho \neq 0$ i.e. ages of husbands and wives seem to be linearly related at the time of marriage.

Example 8.16: Consider the data in example 8.1 where $r = 0.976$, $n = 10$.

Suppose we now have to test $H_0: \rho = 0$ against $H_1: \rho \neq 0$

The test statistic,

$$t = \frac{r - 0}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.976}{\sqrt{\frac{1-(.976)^2}{10-2}}} = 13.57$$



t distribution graph

At 5% level of significance, two tailed table value of t with degree of freedom 8 is 2.306

Decision:

Calculated t (13.57) is greater than the tabled t (2.306). Thus there is no evidence of accepting null hypothesis.

Conclusion:

So it can be concluded that the level of positive correlation is statistically significant.

Example 8.17: Suppose the correlation coefficient between intake of fat and level of cholesterol calculated on the basis of a sample of 10 individuals is $r = 0.913$. Test the hypothesis for population correlation coefficient.

Solution:

Test the hypothesis:

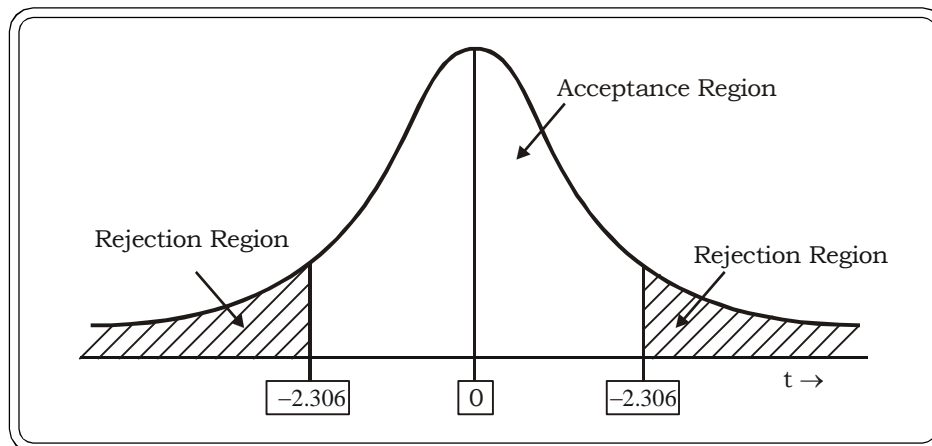
$H_0: \rho = 0$ i.e. there is no linear relationship between intake of fat and level of cholesterol.

$H_1: \rho \neq 0$ i.e. there is significant linear relationship between intake of fat and level of cholesterol.

The test statistic is as follows:

$$t = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n - 2}}} = 6.33$$

At 5% level of significance, the value of t with $(n - 2)$ degrees of freedom is 2.306.



Graph of t-distribution

We reject the null hypothesis and conclude that there is significant linear relationship between intake of fat and the level of cholesterol.

Example 8.18: For a tri variate distribution $r_{12.3} = 0.2425$. For $n = 30$, test the significance of this first order partial correlation coefficient.

Solution:

The null and alternative hypothesis:

$$H_0 : \rho_{12.3} = 0$$

$$H_1 : \rho_{12.3} \neq 0$$

The test statistic

$$t = \frac{r_{12.3}}{\sqrt{1-r_{12.3}^2}} \sqrt{n-k-2} \sim t_{n-k-2}$$

$$\text{Given } r_{12.3} = 0.2425$$

$$n = 30$$

$$k = 1 \quad (\text{no. of secondary subscripts})$$

$$\begin{aligned} \therefore t &= \frac{0.2425}{\sqrt{1-(0.2425)^2}} \sqrt{30-1-2} \\ &= 1.2990 \end{aligned}$$

$$\text{Tabulated } t \left(\frac{.05}{2}, 29 \right) = 2.045$$

Decision

Since calculated $t <$ tabulated t , we may accept the null hypothesis.

Conclusion:

Population partial correlation coefficient is not significant.

Example 8.19: Given $r_{12} = 0.367$, $r_{13} = 0.684$ and $r_{23} = 0.321$. Find $R_{1.23}$ and test its significance when $n = 30$.

Solution:

$$R_{1.23} = \left(\frac{r_{12}^2 + r_{13}^2 - 2r_{12} r_{13} r_{23}}{1 - r_{23}^2} \right)^{\frac{1}{2}}$$

$$r_{12} = 0.367, \quad r_{13} = 0.684, \quad R_{23} = 0.321, \quad n = 30$$

$$\therefore R_{1.23}^2 = \frac{0.4413}{0.8969} = 0.4920$$

$$\text{and } R_{1.23} = 0.7014$$

To test the significance of $R_{1.23}$.

Null and alternative hypothesis

$$H_0 : \rho_{12.3} = 0$$

$$H_1 : \rho_{12.3} \neq 0$$

The test statistic:

$$F = \frac{R_{1,23}^2}{1 - R_{1,23}^2} \frac{n - k - 1}{k} \sim F(k, n - k - 1)$$

Thus,

$$\begin{aligned} F &= \frac{0.4920}{0.5079} \times \frac{30 - 2 - 1}{2} \\ &= 0.9688 \times 13.5 \\ &= 13.0788 \end{aligned}$$

Tabulated $F(2, 27, 0.05) = 3.35$

Decision and Conclusion

Since calculated $F >$ tabulated F , we may reject the null hypothesis and conclude that population multiple correlation coefficient is significant.

8.2 REGRESSION ANALYSIS



This section introduces the concept of regression. Regression analysis summarizes on an average the relationship between two or more variables. Here we discuss the theory behind fitting a regression line, present an algebraic exposition of the ordinary least squares (OLS) regression coefficients, and show several ways to have Excel report regression results dedicated to even more powerful and sophisticated applications of the method of regression.

Regression analysis is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another—the effect of a price increase upon demand, for example, or the effect of changes in the money supply upon the inflation rate. To explore such issues, the investigator assembles data on the underlying variables of interest and employs regression to estimate the quantitative effect of the causal variables upon the variable that they influence. The investigator also typically assesses the “statistical significance” of the estimated relationships, that is, the degree of confidence that the true relationship is close to the estimated relationship.

Regression techniques have long been central to the field of economic statistics (“econometrics”).

8.2.1 Simple Linear Regression

The goal of regression analysis is to determine the values of parameters for a function that cause the function to best fit a set of data observations. In the linear regression model, the dependent variable is assumed to be a linear function of one or more independent variables plus an error introduced to account for all other factors:

As an illustration, suppose that we wish to identify and quantify the factors that determine earnings in the labor market. A moment’s reflection suggests a myriad of factors that are associated with variations in earnings across individuals like: occupation, age, experience, educational attainment, motivation, and innate ability, along with other factors such as race and gender that can be of particular concern. For the time being, suppose we restrict attention to a single factor i.e. education. Regression analysis with a single explanatory variable education in this case is termed “simple regression.”

Let us assume we have a set of data, (x_i, y_i) . If we have reason to believe that there exists a linear relationship between the variables x and y , we can plot the data and draw a “best-fit” straight line through the data. Of course, this relationship is governed by the familiar equation $y = mx + b$. We can then find the slope, m , and y -intercept, b , for the data, which are shown in the figure below.

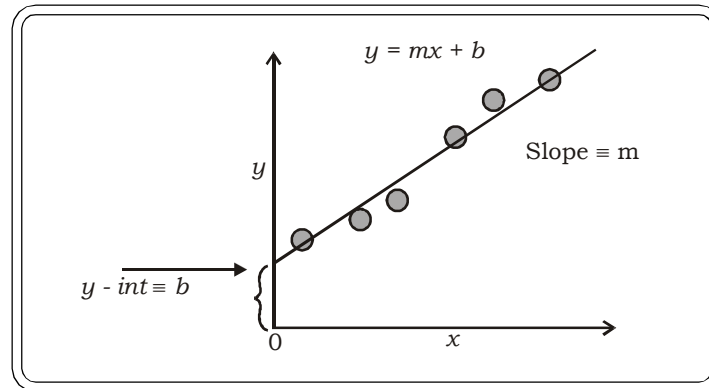


Figure 8.5

Simple Linear Regression

In the above regression equation, y is the *dependent variable* and x is the *independent* or *explanatory variable*. The goal of regression analysis is to obtain estimates of the unknown parameters a and b , which indicate how a change in the independent variable affects the value of the dependent variable.

In economics, the dependent variable might be a family’s consumption expenditure and the independent variables might be the family’s income, number of children in the family, and other factors that would affect the family’s consumption patterns. In political science, the dependent variable might be a state’s level of welfare spending and the independent variables, measures of public opinion and institutional variables that would cause the state to have higher or lower levels of welfare spending. In sociology, the dependent variable might be a measure of the social status of various occupations and the independent variable may be characteristics of the occupations (pay, qualifications, etc.). In psychology, the dependent variable might be individual’s racial tolerance as measured on a standard scale and with indicators of social background as independent variables. In education, the dependent variable might be a student’s score on an achievement test and the independent variables characteristics of the student’s family, teachers, or school.

8.2.1.1 Regression Equation and Regression Coefficients

There are two linear regression equations for two variables say X and Y . One is regression equation of Y on X and the other is regression equation of X on Y . When X is independent and Y is the dependent variable we have regression equation of Y on X . On the other hand if X is dependent and Y is independent we have regression equation of X on Y . The two equation can be written as follows:

(i) Regression equation of Y on X :

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

(ii) Regression equation of X on Y :

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

b_{yx} and b_{xy} are called the regression coefficients and can be defined as follows:

(i) Regression coefficient of Y on X:

$$b_{YX} = \frac{\text{cov}(X, Y)}{\sigma_X^2} = r \frac{\sigma_Y}{\sigma_X} = \frac{\sum xy}{\sum x^2}$$

(ii) Regression coefficient of X on Y:

$$b_{XY} = \frac{\text{cov}(X, Y)}{\sigma_Y^2} = r \frac{\sigma_X}{\sigma_Y} = \frac{\sum xy}{\sum y^2}$$

where $x = X - \bar{X}$ and $y = Y - \bar{Y}$

As defined earlier, $\text{cov}(X, Y)$, r , σ_X , σ_Y are the covariance of X and Y, correlation coefficient, standard deviation of X and standard deviation of Y respectively.

8.2.1.2 Properties of Regression Coefficients

1. The sign of both the regression coefficients are always same.
2. Both the regression coefficients can not simultaneously exceed one. If one is greater than 1, then the other is bound to be less than 1.
3. Regression coefficients are independent of change in origin but not of scale.
4. The correlation coefficient $r = \pm \sqrt{b_{XY} b_{YX}}$
i.e. correlation coefficient is the geometric mean of the regression coefficients.
5. The sign of the correlation coefficient is the same as that of the two-regression coefficients.
6. If $r = 0$ the regression lines are perpendicular to each other.
7. In case of perfect correlation i.e. if $r = \pm 1$, the two regression lines coincide.
8. Both the regression lines pass through (\bar{X}, \bar{Y}) .

Example: 8.20: Calculate regression lines from the following data and estimate x when y is 26 and y when x is 35.

X	10	12	13	17	18	20	24	30
Y	5	6	7	9	13	15	20	21

Solution:

X	Y	x = X - \bar{X}	y = Y - \bar{Y}	xy	x²	y²
10	5	-8	-7	56	64	49
12	6	-6	-6	36	36	36
13	7	-5	-5	25	25	25
17	9	-1	-3	3	1	9
18	13	0	1	0	0	1
20	15	2	3	6	4	9
24	20	6	8	48	36	64
30	21	12	9	108	144	81
144	96			282	310	274

$$\bar{x} = \frac{\sum x}{n} = \frac{144}{8} = 18$$

$$\bar{y} = \frac{\sum y}{n} = \frac{96}{8} = 12$$

Here we have to calculate the two-regression equations.

Regression equation at Y on X is:

$$(y - \bar{y}) = b_{yx}(x - \bar{x})$$

and regression equation of X on Y is:

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$

Regression coefficient of Y on X:

$$b_{yx} = \frac{\sum xy}{\sum x^2} = \frac{282}{310} = 0.91$$

Regression coefficient of X on Y:

$$b_{xy} = \frac{\sum xy}{\sum y^2} = \frac{282}{274} = 1.03$$

Thus regression equation of Y on X will be

$$Y - 12 = 0.91(X - 18)$$

$$Y = 0.91x + 12 - 0.91$$

$$\text{Or } Y = 11.09 + 0.91x$$

And the regression equation X on Y will be

$$(X - 18) = 1.03 (Y - 12)$$

$$\text{or } X = 1.03 Y + 18 - 12.36$$

$$\text{or } X = 5.64 + 1.03 Y$$

Now when $Y = 26$, then

X is estimated from the regression equation of X on Y.

$$\text{Then } X = 5.64 + 1.03 (26)$$

$$\text{Or } X = 32.42$$

Now when $X = 35$

Y is estimated using the regression equation of Y on X.

$$\text{Then } Y = 11.09 + 0.91 (35)$$

$$Y = 42.94$$

Example: 8.21: The following results were obtained in the analysis of data on dry bark in ounces (Y) and age in year (X) of 200 plants:

	X	Y
Average	9.2	16.5
Standard Deviation	2.1	4.2

Correlation coefficient = 0.84

Construct the two lines of regression and estimate the yield of dry bark of a plant of age 8 years. (Patna Univ., B.Sc, 1991)

Solution:

$$\bar{X} = 9.2$$

$$\bar{Y} = 16.5$$

$$\sigma_x = 2.1$$

$$\sigma_y = 4.2$$

Regression line of Y and X

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} = 0.84 \frac{4.2}{2.1} = 1.68$$

Thus

$$Y - 16.5 = 16.8 (X - 9.2)$$

$$Y = 16.8X - 15.456 + 16.5$$

$$Y = 1.68 + 1.044$$

Regression line of X on Y:

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$b_{xy} = r \frac{\sigma_x}{\sigma_y}$$

$$= 0.84 \frac{2.1}{4.2}$$

$$= \frac{1.764}{4.2} = 0.42$$

The regression line of X on Y is:

$$X - 9.2 = 0.42(Y - 16.5)$$

$$X = 0.42Y - 6.93 + 9.2$$

$$X = 0.42Y + 2.27$$

We now have to estimate the yield of dry bark (Y) when age (X) is 8 years. So we consider the regression line of Y on X

$$\text{i.e. } Y = 1.68X + 1.044$$

Putting X = 8

$$Y = 14.484$$

i.e. yield of dry bark when age of the tree is 8 years is 14.484 ounces.

Example 8.22: We are given the following information about advertising expenditure and sales.

	Advertising Expenditure (X) (Rs. Lakhs)	Sales (Y) (Rs. Lakhs)
Mean	10	90
Standard Deviation	3	12

Correlation coefficient = 0.8.

What should be the advertising budget if the company wants to attain sales target of Rs.120 lakhs. (DU, MCA, 1990)

Solution:

Since the company wants to estimate the advertising budget (X) to achieve a sales target of Rs.120 lakhs (Y = 120) we need the regression line of X on Y.

$$(X - \bar{X}) = b_{xy}(Y - \bar{Y})$$

Now,

$$b_{xy} = r \frac{\sigma_x}{\sigma_y}$$

Given $\bar{x} = 10$, $\bar{y} = 90$, $\sigma_x = 3$, $\sigma_y = 12$, $r = 0.8$

Thus

$$b_{xy} = 0.8 \frac{3}{12} = 0.2$$

Thus, regression line of X on Y

$$X - 10 = 0.2 (Y - 90)$$

$$X = 0.2Y - 18 + 10$$

$$X = 0.2Y - 8$$

Thus when $Y = 120$

$$X = 16$$

The advertising budget of the company should be Rs. 16 lakhs to obtain a sales target of Rs.120 lakhs.

Example 8.23: For two variables X and Y, $\bar{X} = 3$, $\bar{Y} = 4$ and $r_{xy} = 0.4$. The line of regression of Y on X is parallel the line $Y = X$. Find the two lines of regression and estimate the mean of X when $\bar{Y} = 1$.

Solution:

Given $\bar{X} = 3$, $\bar{Y} = 4$, $r_{xy} = 0.4$

Line of regression of X on Y:

$$X - \bar{X} = b_{xy} (Y - \bar{Y})$$

Line of regression of Y on X:

$$Y - \bar{Y} = b_{yx} (X - \bar{X})$$

$$b_{xy} = r_{xy} \frac{\sigma_x}{\sigma_y} \quad \text{and} \quad b_{yx} = r_{yx} \frac{\sigma_y}{\sigma_x}$$

Since the line of regression of Y on X is parallel to the line $Y = X$, slope of both lines are same
 $\Rightarrow b_{yx} = 1$

$$\text{Thus } \frac{\sigma_y}{\sigma_x} = \frac{b_{yx}}{r_{xy}} = \frac{1}{0.4} = 2.5$$

$$\text{and } \frac{\sigma_x}{\sigma_y} = \frac{1}{2.5} = 0.4$$

$$\text{and } b_{xy} = r_{xy} \frac{\sigma_x}{\sigma_y} = (0.4)(0.4) = 0.16$$

Thus, line of regression of X on Y:

$$\begin{aligned} X - 3 &= 0.16(Y - 4) \\ \Rightarrow X &= 0.16Y - 0.64 + 3 \\ \Rightarrow X &= 0.16Y - 2.36 \end{aligned}$$

Line of regression of Y on X:

$$\begin{aligned} Y - 4 &= 1(x - 3) \\ \Rightarrow Y &= X + 1 \end{aligned}$$

For estimating the mean of X, we will use the line of regression of X on Y i.e.

$$X = 0.16Y + 2.36$$

Since the regression lines pass through (\bar{X}, \bar{Y}) ,

$$\therefore \bar{X} = 0.16 \bar{Y} + 2.36$$

When $\bar{Y} = 1$,

$$\bar{X} = 2.52$$

Example 8.24: The lines of regression for a bivariate distribution are given by

$$X + 9Y = 7 \quad \text{and} \quad Y + 4X = \frac{49}{3}$$

(i) Calculate the correlation coefficient.

(ii) \bar{X} and \bar{Y} .

Solution:

Consider

$$X = -9Y + 7 \quad \text{(Line of X on Y)}$$

$$Y = -4X + \frac{49}{3} \quad \text{(Line of Y on X)}$$

$$b_{xy} = -9, \quad b_{yx} = -4$$

$$\Rightarrow r \frac{\sigma_x}{\sigma_y} = -9 \quad \text{and} \quad r \frac{\sigma_y}{\sigma_x} = -4$$

$$\Rightarrow \frac{\sigma_x}{\sigma_y} = \frac{-9}{r} \quad \text{and} \quad \frac{\sigma_x}{\sigma_y} = \frac{-r}{4}$$

From the two equations above,

$$\frac{9}{r} = \frac{r}{4} \Rightarrow r^2 = 36 \Rightarrow r = \pm 6, \text{ which is not possible because } -1 \leq r \leq 1$$

Therefore our assumption about the lines of regression must be wrong.

Let us now assume,

$$Y = -4X + \frac{49}{3} \text{ as the line of Regression of X on Y}$$

$$\Rightarrow 4X = -4Y + \frac{49}{3}$$

$$\Rightarrow X = -0.25Y + 4.08$$

Let line of Y on X be,

$$X = -9Y + 7$$

$$\Rightarrow 9Y = -X + 7$$

$$Y = -0.11 + 0.78$$

$$\text{thus, } b_{xy} = -0.25$$

$$\Rightarrow r \frac{\sigma_x}{\sigma_y} = -0.25$$

$$\Rightarrow \frac{\sigma_x}{\sigma_y} = -\frac{0.25}{r} \quad \dots (1)$$

$$\text{And } b_{yx} = -0.11$$

$$\frac{\sigma_y}{\sigma_x} = \frac{-0.11}{r} \Rightarrow \frac{\sigma_x}{\sigma_y} = \frac{-r}{0.11} \quad \dots (2)$$

Comparing (1) and (2),

$$\frac{-0.25}{r} = \frac{-r}{0.11}$$

$$\Rightarrow r^2 = 0.02$$

$$\Rightarrow r = 0.16$$

and since r assumes the sign of the regression coefficient,

$$r = -0.16$$

(ii) Since both the regression lines pass through (\bar{X}, \bar{Y}) , we only need to solve the equations to obtain \bar{X} and \bar{Y} .

$$\bar{X} = -9\bar{Y} + 7 \quad \dots (1)$$

$$+4\bar{X} = -\bar{Y} + \frac{49}{3} \quad \dots (2)$$

$$\bar{X} = -9\bar{Y} + 7$$

$$36\bar{X} = -9\bar{Y} + 147$$

$$-35\bar{X} = -140$$

$$\Rightarrow \bar{X} = 4$$

$$\begin{aligned}\therefore \bar{Y} &= -4(4) + \frac{49}{3} \\ &= -16 + 16.33 \\ &= -0.33\end{aligned}$$

Thus $(\bar{X}, \bar{Y}) = (4, -0.33)$

8.2.1.3 Least Square Method and Regression Equation

The method of least squares assumes that the best-fit curve of a given type is the curve that has the minimal sum of the deviations squared (*least square error*) from a given set of data. When we are using the relation $Y = a + bX$, it indicates the exact relation. In the real world however, the relationship is not exact. In the following figure 8.6, the dotted points are the actual data, whereas the regression line is showing the straight line relation fitted between X and Y.

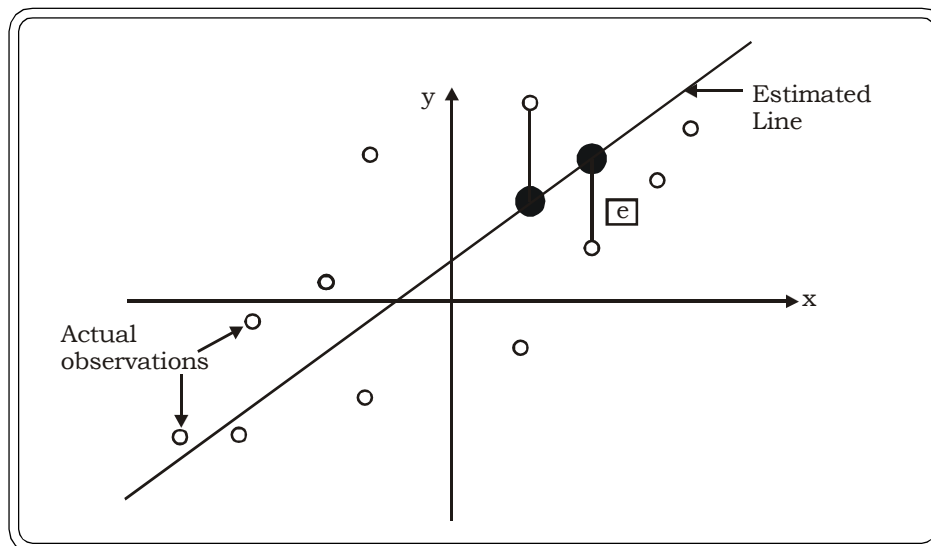


Figure 8.6

Line of Best Fit

A regression line can be drawn through these scatter points. But the question arises, which one would be the best? Intuitively, the best line would be the line where the departure of the regression line from the scatter points are less. The differences between the actual values and the estimated values should be the minimum or least. However, if only the deviations are taken, the ultimate sum will be zero because some deviations will be positive and some are negative. To overcome the problem, the square of the deviations was proposed. This is the background of the least square method.

Consider the data points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, where x is the independent variable and y is the dependent variable. Let $y = f(x)$ be the fitted curve. $f(x)$ has the deviation (error) d from each data point, i.e., $d_1 = y_1 - f(x_1), d_2 = y_2 - f(x_2), \dots, d_n = y_n - f(x_n)$. These differences are termed as errors. According to the method of least squares, the best fitting curve has the property that the sum of square of these errors is minimum.

Thus

$$D = d_1^2 + d_2^2 + \dots + d_n^2 = \sum_{i=1}^n d_i^2 = \sum_{i=1}^n [y_i - f(x_i)]^2 \text{ must be minimum.}$$

$$\text{i.e. } D = \sum_{i=1}^n [y_i - f(x_i)]^2 = \sum_{i=1}^n [y_i - (a + bx_i)]^2 \text{ must be minimum.}$$

To minimize this function, we differentiate it with respect to the two unknowns a and b , using the maxima and minima principles of differential calculus, as follows:

$$\begin{cases} \frac{\partial D}{\partial a} = 2 \sum_{i=1}^n [y_i - (a + bx_i)] = 0 \\ \frac{\partial D}{\partial b} = 2 \sum_{i=1}^n x_i [y_i - (a + bx_i)] = 0 \end{cases}$$

Expanding the above equations, we have:

$$\begin{cases} \sum_{i=1}^n y_i = a \sum_{i=1}^n 1 + b \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i = a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 \end{cases}$$

These two equations are known as normal equations. The unknown coefficients a and b can be obtained by solving these normal equations:

$$\hat{b} = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$\hat{a} = \bar{y} - b\bar{x}$$

Thus, the estimated regression line of y on x by the method of least squares is $\hat{y} = \hat{a} + \hat{b}x$

Example: 8.25: A manufacturing company is interested in evaluating the annual sales of the company in lakhs of Rs over the past 11 years. For this, they have compiled the data related to the annual sales of the company for the past 11 years. Relate annual sales to the years and predict the sales for 2006.

Year (x)	Annual sale(Lakhs of Rupees)
1995	1
1996	5
1997	4
1998	7
1999	10
2000	8
2001	9
2002	13
2003	14
2004	13
2005	18

Solution:

In the above problem, the independent variable is the year and the dependent variable is the annual sales.

As the years do not have any numerical meaning, we can adjust as follows:

Year	Y	X	X ²	XY
1995	1	-5	25	-5
1996	5	-4	16	-20
1997	4	-3	9	-12
1998	7	-2	4	-14
1999	10	-1	1	-10
2000	8	0	0	0
2001	9	1	1	9
2002	13	2	4	26
2003	14	3	9	42
2004	13	4	16	52
2005	18	5	25	90
	$\Sigma Y=102$	$\Sigma X=0$	$\Sigma X^2=110$	$\Sigma XY=158$

Here, we have chosen the origin at the middle year of 2000 where the value of X, the actual independent variable, is 0. X values for other years are calculated accordingly.

Let the straight line equation be $Y = a + bX$, where a and b are the two parameters and can be calculated by following the formulae:

$$b = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2}$$

$$a = \bar{Y} - b\bar{X}$$

n = number of observations = 11

Putting the values from the table we have

$$\bar{Y} = \frac{102}{11} = 9.27 \quad \bar{X} = 0$$

$$\text{Thus } b = \frac{11 \times 158 - 0 \times 102}{11 \times 110 - 0^2} = 1.44$$

$$a = 9.27$$

Thus the fitted straight line is $Y = 9.27 + 1.44X$

During 2006, X will be 6

Then Y will be

$$9.27 + 1.44 \times 6 = 17.91$$

Thus, the above regression analysis predicts the sale of 2006 as 17.91 Lakhs of Rupees.

Example 8.26: Personal Manager of a large industrial unit is interested to find a measure that can be used to fix the wages (yearly) of skilled workers. On experimental basis, the data on the length of service and their yearly wages (in Rs.'000) from a group of 10 randomly selected skilled workers are given below:

Length of Service (X)	11	7	9	5	8	6	10	12	3	4
Yearly Wages	14	11	10	9	13	10	14	16	6	7

- (i) Develop the regression equation of wage (Y) on the length of service X.
- (ii) On the basis of (i) what initial pay the personnel manager should give to a skilled worker who has put in thirteen years of service on a similar basis, in another industry.

Solution: (i)

X	Y	X²	XY
11	14	121	154
7	11	49	77
9	10	81	90
5	9	25	45
8	13	64	104
6	10	36	60
10	14	100	140
12	16	144	192
3	6	9	18
4	7	16	28
75	110	645	908

The regression equation is:

Wage (Y) = a + b x length of service (X)

From the above table,

$$\Sigma X = 75, \Sigma Y = 110, \Sigma X^2 = 645, \Sigma XY = 908,$$

$$n = 10$$

Constants a and b are estimated as follows:

$$\begin{aligned} \hat{b} &= \frac{n \Sigma XY - \Sigma X \Sigma Y}{n \Sigma X^2 - (\Sigma X)^2} \\ &= \frac{10(908) - (75)(110)}{10(645) - (75)^2} = \frac{9080 - 8250}{6450 - 5625} \\ &= \frac{830}{825} = 1.006 \end{aligned}$$

$$\begin{aligned} a &= \bar{Y} - b\bar{X} \\ &= 11 - (1.006) 7.5 \\ &= 11 - 7.545 = 3.455 \end{aligned}$$

Thus, the regression equation is

$$Y = 3.455 + 1.006X$$

(ii) Salary of a worker who has 13 years of service in a similar industry is obtained by putting $X = 13$ in this equation.

$$\begin{aligned} &\text{Thus } 3.455 + (1.066) 13 \\ &= 16.533 \end{aligned}$$

i.e. approximately Rs.16, 500

Example 8.27: The production manager of a firm is interested in studying the relationship between intelligence & productivity of the worker. 10 workers are selected at random and their score on aptitude test & productivity indices are compiled in a tabular form as follows. Express this relation with a suitable linear regression model and predict if the aptitude score is 70, what would be the productivity of the worker.

Aptitude Score	78	65	78	48	53	70	72	65	60	67
Productivity Index	79	60	62	40	52	80	85	62	68	60

Solution:

Let aptitude score be the independent variable x .

Productivity be the dependent variable y .

So the regression equation will be:

$$y = a + bx$$

The least square estimate of a and b can be calculated by using the following formulae:

$$\hat{b} = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$a = \bar{y} - \hat{b}\bar{x}$$

x	y	x^2	xy
78	79	6084	6162
65	60	4225	3900
73	62	5329	4526
48	40	2309	1920
53	52	2809	2756
70	80	4900	5600
72	85	5184	6120
65	62	4225	4030
60	68	3600	4080
67	60	3721	3660

$$n = 10, \Sigma x = 645, \Sigma y = 648, \Sigma x^2 = 42381, \Sigma xy = 42754$$

$$\text{Now } b = \frac{10 \times 42754 - 645 \times 648}{10 \times 42381 - (645)^2} = \frac{427540 - 417960}{423810 - 416025} = \frac{9580}{7785} = 1.23$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = \frac{648}{10} - 1.23 \frac{645}{10} = 64.8 - 79.34 = -14.54$$

Thus the regression equation will be

$$y = -14.54 + 1.23x$$

Now, the productivity when the aptitude score is 70 is,

$$\begin{aligned} y &= -14.54 + 1.23 \times 70 \\ &= -14.54 + 86.1 \\ &= 71.56 \end{aligned}$$

Example 8.28: In 2005 a researcher collected data on saving and investment from 16 households. Household savings had a mean of Rs.6565.00 with a variance of Rs.250.00. As against this, mean investment was Rs.4525.00 with variance of Rs.520.00. If the coefficient of correlation between saving and investment is 0.67. What would the value of saving if investment is Rs.9000.00?

Solution:

Let x be the investment i.e. the independent variable and y be the saving i.e. dependent variable. In this case, as the data suggests, we have to use the regression equation of the following form.

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

Given:

$$\bar{y} = 6565$$

$$\sigma_y^2 = 250$$

$$\bar{x} = 4525$$

$$\sigma_x^2 = 520$$

$$r = 0.67$$

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} = 0.67 \frac{15.81}{22.80} = 0.67 \times 0.69 = 0.46$$

Thus, the regression equation of y on x becomes:

$$y - 6565 = 0.46(x - 4525)$$

$$\text{Or } y = 0.46x + 6565 - 2081.5$$

$$\text{Or } y = 0.46x + 4483.5$$

Now when $x = 9000$,

$$\begin{aligned} y &= 0.46(9000) + 4483.5 \\ &= 8623.5 \end{aligned}$$

If the investment is of Rs.9000 the saving will be Rs. 8623.5

Example 8.29: From a regression analysis of input output, data the following results have been calculated.

$$5y = 4x + 33 \quad (\text{Regression line of } y \text{ on } x)$$

$$20x = 9y + 107 \quad (\text{Regression line of } x \text{ on } y)$$

$$\sigma_x = 3$$

Find out the coefficient of correlation, standard deviation of y and (\bar{x}, \bar{y}) .

Solution:

The regression equation of y on x is

$$5y = 4x + 33 \quad \dots(1)$$

and that of x on y is:

$$20x = 9y + 107 \quad \dots(2)$$

From equation (1) we can have regression coefficient of y on x , b_{yx} and similarly by from equation (2) b_{xy} :

$$5y = 4x + 33$$

$$\text{or } y = \frac{4}{5}x + \frac{33}{5}$$

$$\Rightarrow b_{yx} = \frac{4}{5}$$

$$20x = 9y + 107$$

$$\Rightarrow b_{xy} = \frac{9}{20}$$

From the relationship between r and regression coefficients we have:

$$\begin{aligned} r &= \sqrt{b_{yx} \cdot b_{xy}} \\ &= \sqrt{\frac{4}{5} \cdot \frac{9}{20}} = 0.6 \end{aligned}$$

Thus, the correlation coefficient is 0.6.

Again,

$$b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$$

$$\sigma_y = \frac{\sigma_x \cdot b_{yx}}{r} = \frac{3 \times \frac{4}{5}}{0.6} = \frac{\frac{12}{5}}{0.6} = \frac{2.4}{0.6} = 4$$

Thus, the standard deviation of $y = 4$.

The two regression equations are given here simplifying we can get,

$$y = \frac{4}{5}x + \frac{33}{5}$$

$$x = \frac{9}{20}y + \frac{107}{20}$$

Both the equations will be satisfied by the mean values \bar{x} and \bar{y} . So we can write

$$\bar{y} = \frac{4}{5}\bar{x} + \frac{33}{5} \text{ and}$$

$$\bar{x} = \frac{9}{20}\bar{y} + \frac{107}{20}$$

Solving these two equations we will be getting the means as:

$$\bar{x} = 13$$

$$\bar{y} = 17$$

Example 8.30: For a pen manufacturing company, the daily production of pen and the number of workers assigned is given in the following table. If 10 workers are employed, calculate how many pens would be produced.

No. of workers	No. of pens produced in one day
6	350
7	370
7	390
8	400
8	410
6	324
6	340
7	399
9	500
9	500

Solution:

Let workers be independent variable x and number of pens produced be the dependent variable y .

The linear regression equation of y on x will be $y = a + bx$.

The estimates are $b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$

and $a = \bar{y} - \hat{b}\bar{x}$

$n = 10$

No. of workers (x)	No. of pens produced in one day (y)	x^2	xy
6	350	36	2100
7	370	49	2590
7	390	49	2730
8	400	64	3200
8	410	64	3280
6	324	36	1944
6	340	36	2040
7	399	49	2793
9	500	81	4500
9	500	81	4500
$\sum x = 73$	$\sum y = 3983$	545	29677

$$\hat{b} = \frac{10 \times 29677 - 73 \times 3983}{10 \times 545 - (73)^2}$$

$$= \frac{296770 - 290759}{5450 - 3329} = \frac{6011}{121} = 49.68$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$= \frac{3983}{10} - 49.68 \frac{73}{10}$$

$$= 398.3 - 362.66 = 35.64$$

The estimated regression equation is:

$$Y = 35.64 + 49.68x$$

Now, when $x = 10$

$$Y = 35.64 + 496.8 = 532$$

If there are 10 workers, then the estimated number of pens produced will be 532.

Example 8.31: An industrial engineer collected the following data on experience and performance rating of 8 operators.

Operators	1	2	3	4	5	6	7	8
Experience (year)	16	12	18	4	3	10	5	12
Performance rating	87	88	89	68	58	80	70	85

- Does the data give evidence that experience improves performance?
- Estimate the performance rating of an operator who has 9 and 15 years of experience?

Solution:

Let years of experience be the independent variable x and performance rating be the dependent variable y .

The regression equation at y on x is

$$y = a + bx$$

The estimates of b and a can be calculated by the following the least square formula:

$$\hat{b} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

x	y	x^2	xy
16	87	256	1392
12	88	144	1056
18	89	324	1602
4	68	16	272
3	58	9	174
10	80	100	800
5	70	25	350
12	85	144	1020
$\sum x = 80$	$\sum y = 625$	1018	6666

$$\begin{aligned} \hat{b} &= \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \\ &= \frac{8 \times 6666 - 80 \times 625}{8 \times 1018 - (80)^2} \end{aligned}$$

$$= \frac{53328 - 50000}{8144 - 6400} = \frac{3328}{1744} = 1.91$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$= \frac{625}{8} - 1.91 \frac{80}{8}$$

$$= \frac{625 - 152.8}{8} = 59.03$$

Thus, the regression equation will be

$$y = 59.03 + 1.91 x$$

1. \hat{b} is positive. Thus it can be said that there is a positive direct relationship between the years of experience and performance. Thus, from the above data it can be said that experience improves performance.
2. Now, when $x = 9$ and $x = 15$, we have to calculate the value of y .

At $x = 9$

$$\begin{aligned} y &= 59.03 + 1.91 \times 9 \\ &= 59.03 + 17.19 = 76.22 = 76 \end{aligned}$$

At $x = 15$

$$\begin{aligned} Y &= 59.03 + 1.91 \times 15 \\ &= 87.68 = 88 \end{aligned}$$

Thus, the estimated performance rating will be 77 and 88 respectively with the year of experiences as 9 and 15 years.

Example 8.32: The data on annual turnover and the no. of staff for the last 8 years of a manufacturing company is given in the following table.

Years	Business turn over (Rs. Crores)	No. of Staff
1998	45	2600
1999	50	3000
2000	60	3100
2001	75	3530
2002	80	3850
2003	110	4300
2004	150	5870
2005	170	7150

Relate the business turnover and number of staff of the company. Now suppose the company is targeting a turnover of Rs.200 crores. How many new staff should be employed? Give an estimate.

Solution:

Let the turnover be the independent variable x and no. of staff be the dependent variable y .

We are relating these two variables by a simple linear regression model as follows:

$$y = a + bx$$

b can be estimated as

$$\hat{b} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

and that of a as

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

Years	x	y	x^2	xy
1998	45	2600	2025	117000
1999	50	3000	2500	150000
2000	60	3100	3600	186000
2001	75	3530	5625	264750
2002	80	3850	6400	308000
2003	110	4300	12100	473000
2004	150	5870	22500	880500
2005	170	7150	28900	1215500
	$\sum x = 740$	$\sum y = 33400$	$\sum x^2 = 83650$	$\sum xy = 3594750$

$$\hat{b} = \frac{8 \times 3594750 - 740 \times 33400}{8 \times 83650 - (740)^2} = \frac{4042000}{121600} = 33.24$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$= \frac{33400}{8} - 33.24 \frac{740}{8} = 1100.3$$

Thus, the linear regression equation that shows the relationship between the annual turnover and the number of staff employed in the given company.

$$y = 1100.3 + 33.24x$$

The company is targeting the annual turnover of Rs.200 crores i.e., with $x = 200$ what will be y .

$$\begin{aligned} y &= 1100.3 + 33.24 \times 200 = 7748.3 \\ &= 7748 \end{aligned}$$

Thus, the estimated staff should be 7748.

8.2.1.4 Explained and Unexplained Variation

As we have already discussed, the regression line does not explain the total variation of an observation. By the least square method, we are calculating the best fitted line. Some part still remains unexplained.

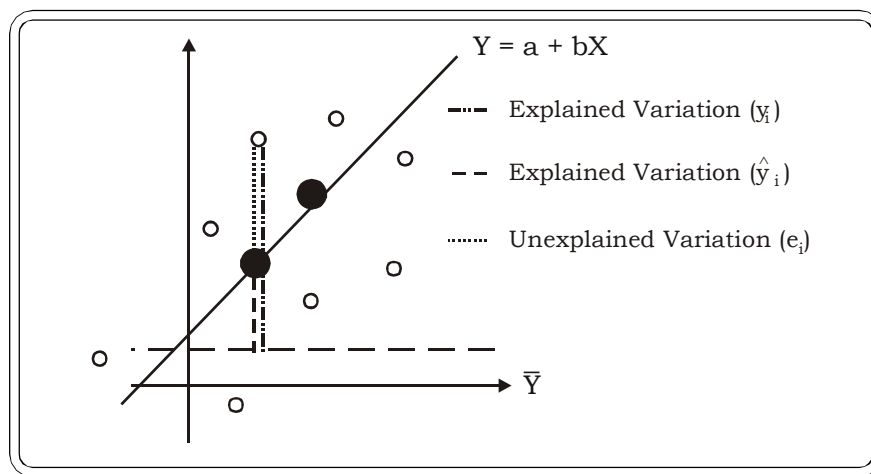


Figure 8.7

Explained and Unexplained Variation

In figure 8.7, we can see that there are two parts of total variation viz. explained and unexplained.

$$\text{Total variation} = \text{Explained variation} + \text{Unexplained variation}$$

If all the points fall on the regression line, which hardly happens in reality, then the unexplained variation will be zero. Total variations, explained variation and unexplained variation of a regression can be calculated by the following formulae:

$$\text{Total Variation} = \sum (Y - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

$$\text{Explained Variation} = \sum (\hat{Y} - \bar{Y})^2 = a \sum Y + b \sum XY - \frac{(\sum Y)^2}{n}$$

$$\text{Unexplained variation} = \sum (Y - \hat{Y})^2 = \sum Y^2 - a \sum Y - b \sum XY$$

where Y denotes the actual observations and

\hat{Y} denotes the estimated values.

The coefficient of determination (r^2) discussed in section 8.1.4 can also be expressed in terms of explained and total variation as follows

$$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{\sum (\hat{Y} - \bar{Y})^2}{\sum (Y - \bar{Y})^2} = 1 - \frac{\text{Unexplained variation}}{\text{Total variation}}$$

Example 8.33: The correlation coefficient between two variables X and Y is 0.8. The unexplained component of variation is given as 50.3. Find what percentage of the variation is explained by the straight-line relationship. Also find the other component of variation.

Solution:

$$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

$$0.8^2 = \frac{\text{Total} - \text{unexplaine d variation}}{\text{Total variation}}$$

$$\Rightarrow 1 - \frac{\text{unexplaine d variation}}{\text{Total variation}} = 0.64$$

$$\Rightarrow \frac{\text{unexplaine d variation}}{\text{Total variation}} = 0.36$$

$$\Rightarrow \text{Total variation} = 139.72$$

Thus, Explained variation = $139.72 - 50.3 = 89.42$

The percentage of variation explained by the straight-line relationship is 64%.

Example 8.34: The following information relates the annual yield of rice per acre (y) to the irrigation (x), $n= 9$,

$$\sum_{i=1}^9 (x_i - \bar{x})^2 = 91.5$$

$$\sum_{i=1}^9 (y_i - \bar{y})^2 = 2000$$

$$\sum_{i=1}^9 (x_i - \bar{x})(y_i - \bar{y}) = 350$$

- (i) Calculate Karl Pearson's correlation coefficient.
- (ii) Calculate Coefficient of Determination and interpret its meaning.

Solution:

$$\begin{aligned}
 \text{(i) } r_{xy} &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \\
 &= \frac{350}{\sqrt{91.5} \sqrt{2000}} \\
 &= \frac{350}{(9.56)(44.72)} = \frac{350}{427.54} \\
 &= 0.82
 \end{aligned}$$

(ii) Co-efficient of Determination = 0.67

This implies 67% of the variation in the yield of rice is due to irrigation.

8.2.1.5 Standard Error of Estimate

The standard error of the estimate is a measure of the accuracy of predictions made with a regression line. Standard error can be symbolized as S_e . The concept of standard error is very much similar to the concept of standard deviation that we have discussed in chapter 4. Standard deviation measures the dispersion of a set of observations about the mean while, standard error of the estimated regression line measures the variability of the scatter from the regression line. Standard error can be estimated by the following formula:

$$S_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$$

where \hat{y} is the estimated y , the dependent variable and, n the number of observations. Here we are losing 2 degrees of freedom because of the two parameters a and b estimated in the regression equation $y = a + bx$. Hence the S_e is calculated by dividing the unexplained variation by $n - 2$.

For simplifying calculations, the formula of S_e can be written as

$$S_e = \sqrt{\frac{\sum y^2 - a \sum y - b \sum xy}{n - 2}}$$

The standard error of the estimate is a measure of the reliability of the estimating equation.

The smaller the value of S_e , the more reliable is the regression equation.

Example 8.35: A general manager at a car rental agency in Delhi is interested to establish a relationship between the car mileage and maintenance cost of the cars. A random sample of 9 cars was selected from the past record and mileage and maintenance cost of these cars is shown in the following table:

Mileage (thousand)	Maintenance cost (Rs.)
15	5.0
10	4.0
12	4.5
11	4.4
9	3.9
7	3.0
8	3.5
3	2.8
2	2.0

Find out the regression equation. Also compute the standard error of the estimate and explained and unexplained variation.

Solution:

Let we consider the mileage as independent variable x and maintenance cost as dependent variable y .

The regression equation is :

$$y = a + bx$$

The estimates of b and a can be calculated by using the following formulae:

$$\hat{b} = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

Mileage (x)	Maintenance Cost (y)	x^2	xy	y^2
15	5.0	225	75	25
10	4.0	100	40	16
12	4.5	144	54	20.25
11	4.4	121	48.4	19.36
9	3.9	81	35.1	15.21
7	3.0	49	21	9
8	3.5	64	28	12.25
3	2.8	9	8.4	7.84
2	2.0	4	4	4
77	33.1	797	313.9	128.91

$$\hat{b} = \frac{9 \times 313.9 - 77 \times 33.1}{9 \times 797 - (77)^2} = \frac{276.4}{1244} = .22$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$\frac{33.1}{9} - .22 \frac{77}{9} = 1.80$$

Thus the estimated regression equation will be:

$$y = 1.80 + 0.22x$$

Next is the calculation of standard error, which is as follows:

$$S_e = \sqrt{\frac{\sum y^2 - a \sum y - b \sum xy}{n-2}}$$

$$\sum y^2 = 128.91$$

$$= \sqrt{\frac{(128.91) - (1.80 \times 33.1) - (0.22 \times 313.9)}{9-2}} = \sqrt{\frac{.27}{7}}$$

$$= 0.196$$

Thus, the calculated standard error is 0.196

$$\begin{aligned} \text{Explained variation} &= a \sum y + b \sum xy - \frac{(\sum y)^2}{n} \\ &= 1.80(33.1) + .22(313.9) - \frac{(33.1)^2}{9} = 6.90 \end{aligned}$$

$$\begin{aligned} \text{Unexplained variation} &= \sum y^2 - a \sum y - b \sum xy \\ &= 128.91 - 0.1.80(33.1) - 0.22(313.9) = 0.27 \end{aligned}$$

Example 8.36: A financial analyst obtained the following information relating to return on security A and that of market portfolio M for the past 8 years.

Year	Return on Security (A)	Market Portfolio (M)
1	10	12
2	15	14
3	18	13
4	14	10
5	16	9
6	16	13
7	18	14
8	4	7

- (i) Develop an estimating equation that best describes these data. Find standard error of estimate.
- (ii) Find the coefficient of determination.
- (iii) Determine the percentage of total variation in security return being explained by the return on the market portfolio. (MFC, DU, 1998)

Solution:

Year	Return on A (x)	Return on M (y)	x ²	y ²	xy
1	10	12	100	144	120
2	15	14	225	196	210
3	18	13	324	169	234
4	14	10	196	100	140
5	16	9	256	81	144
6	16	13	256	169	208
7	18	14	324	196	252
8	4	7	16	49	28
	111	92	1697	1104	1336

Let the estimating equation be

$$y = a + bx$$

a and b are estimated using the least square method:

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$\sum xy = 1336, \sum x = 111, \sum y = 92$$

$$\sum x^2 = 1697, \sum y^2 = 1104, n = 8.$$

Thus,

$$\hat{b} = \frac{8(1336) - (111)(92)}{8(1697) - (111)^2}$$

$$= \frac{10688 - 10212}{13576 - 12321}$$

$$= \frac{476}{1255} = 0.38$$

$$\begin{aligned}
 a &= \bar{y} - b\bar{x} \\
 &= 11.5 - (0.38)(13.875) \\
 &= 11.5 - 5.2725 \\
 &= 6.23
 \end{aligned}$$

Thus the regression equation is

$$y = 6.23 + 0.38x$$

The standard error of estimate

$$\begin{aligned}
 S_e &= \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-2}} \\
 &= \sqrt{\frac{1104 - (6.23)(92) - (0.38)(1336)}{6}} \\
 &= \sqrt{\frac{1104 - 573.16 - 506.68}{6}} = 2.01
 \end{aligned}$$

(ii) Explained variation

$$\begin{aligned}
 &= a\sum y + b\sum xy - \frac{(\sum y)^2}{n} \\
 &= (6.23)(92) + (0.38)(1336) - \frac{(92)^2}{8} \\
 &= 573.16 + 507.68 - 1058 \\
 &= 22.84
 \end{aligned}$$

Unexplained variation

$$\begin{aligned}
 &= \sum y^2 - a\sum y - b\sum xy \\
 &= 1104 - (6.23)(92) - (0.38)(1336) \\
 &= 1104 - 573.16 - 507.68 \\
 &= 23.16
 \end{aligned}$$

Total variation = 46

Thus,

$$\begin{aligned}
 \text{Coefficient of determination} &= \frac{\text{Explained Variation}}{\text{Total Variation}} \\
 &= \frac{22.84}{46} = 0.49
 \end{aligned}$$

(iii) From the value of the coefficient of determination we may conclude that the percentage of total variation in security return being explained by the return on the market portfolio is 49%.

8.2.1.6 Testing the Significance of Regression Coefficients

The least square estimate \hat{b} is obtained from a sample of observations on y and x . Now, as these values are estimated values, it is necessary to apply test of significance in order to measure the significance of the linear relationship.

Consider the fitted simple linear regression line

$$\hat{y} = a + bx$$

For testing a null hypothesis of the form

$$H_0 : \beta = \beta_0 \text{ i.e. the population regression coefficient is } \beta_0$$

against the alternative hypothesis

$$H_1 : \beta \neq \beta_0 \text{ i.e. the population regression coefficient is not equal to } \beta_0,$$

the test statistic is:

$$t = \frac{(b - \beta_0) \sqrt{(n-2) \Sigma(x - \bar{x})^2}}{\sqrt{\Sigma(y - \hat{y})^2}} \sim t_{n-2}$$

where b is the estimated value from the regression line.

\hat{y} are the estimated values.

For calculation purposes, the above value of t can be simplified to

$$t = \frac{(b - \beta_0) \sqrt{(n-2) \Sigma(x - \bar{x})^2}}{\sqrt{\Sigma y^2 - a \Sigma \Sigma - b \Sigma \Sigma x}} = \frac{b - \beta_0}{s_e} \sqrt{\Sigma(x - \bar{x})^2}$$

Decision

For a particular level of significance, H_0 is rejected if calculated $|t| > \text{tabulated } t \left(\frac{\alpha}{2}, n-2 \right)$, else it is accepted.

Example 8.37: For the data in example 8.36, test the hypothesis

$$H_0 : \beta = 0 \text{ (There is no linear relationship)}$$

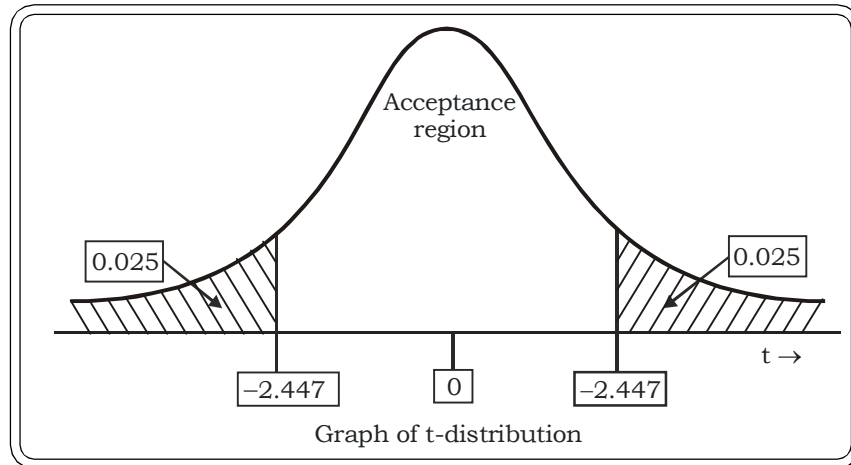
(There is linear relationship)

against $H_1 : \beta \neq 0$ (use 5% level of significance)

Solution:

The test statistic

$$t = \frac{b - \beta}{s_e} \sqrt{\Sigma(x - \bar{x})^2} = \frac{0.38 - 0}{2.01} \times 12.52 = 2.36$$



The tabulated value of t at 5% level of significance and 6 degrees of freedom = 2.447.

Thus, we may accept the null hypothesis. This leads to the conclusion that the linear relationship is not significant.

Example 8.38: The following equation relates to the resale value of a model of car (Y), in hundred of rupees as compared to the number of years (X) it has been in use.

$$Y = 34.68 - 4.22x$$

Test the hypothesis that the annual rate of depreciation is Rs.400.

Given $s_b = 0.232$ and $n = 12$ where $s_b^2 = \frac{s_e^2}{\Sigma(x - \bar{x})^2}$

Solution:

The null hypothesis:

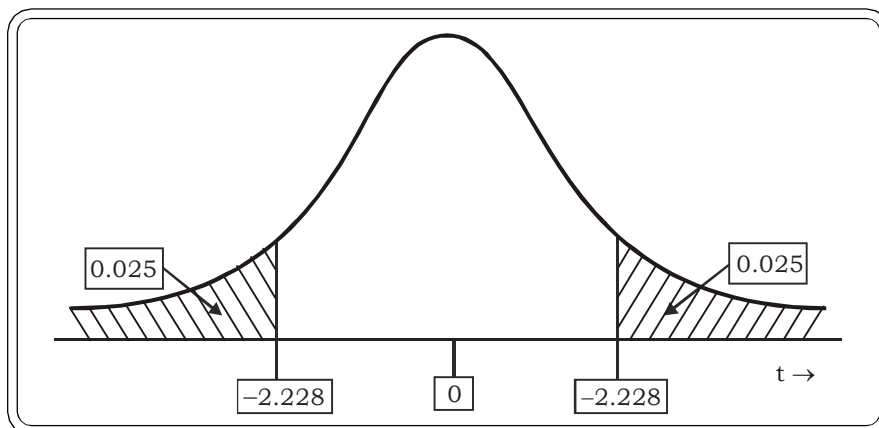
$$H_0: \beta = -4 \text{ i.e. the annual rate of depreciation is Rs.400.}$$

The alternative hypothesis:

$$H_1: \beta \neq -4 \text{ i.e. the annual rate of depreciation is not Rs.400.}$$

The test statistic

$$\begin{aligned} t &= \frac{b - \beta}{s_e} \sqrt{\Sigma(x - \bar{x})^2} \\ &= \frac{-4.22 - (-4)}{0.232} = -0.94 \end{aligned}$$



The tabulated value of t at 5% level of significance and 10 degrees of freedom = 2.228.

Thus, we may accept the null hypothesis and conclude that the annual rate of depreciation is Rs. 400.

8.2.2 Multiple Regression

In section 8.2.1 we have discussed the simple linear regression model, which dealt with the linear relationship between two variables, one dependent and one independent. Multiple regression is used to learn the relationship between more than one independent variable and a single dependent variable. Thus, multiple regression is an extension of simple regression. However, the basic concepts are the same i.e. the prediction of the dependent variable depending on the independent variable(s).

The generalized multiple regression equation can be expressed as follows:

$$Y = B_0 + B_1X_1 + B_2X_2 + \dots + B_nX_n$$

where Y is the dependent variable, X_1, X_2, \dots, X_n are the independent variables, β_i 's are regression coefficients, associated with the i^{th} independent variable, $i = 1, 2, \dots, n$.

8.2.2.1 Multiple Regression with Two Independent Variables

Let us start with two independent variables. It is to be remembered that the same method can be applied for more than two independent variable. The regression equation will be:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2$$

Applying the same least square method as discussed in sub section 8.2.1.3, we will be getting the normal equations first. As we have three coefficients $\beta_0, \beta_1, \beta_2$, there will be three normal equations. These equations are solved to get the values of these parameters.

Normal Equations

$$(1) \sum y = n\beta_0 + \beta_1 \sum x_1 + \beta_2 \sum x_2$$

$$(2) \sum x_1y = \beta_0 \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_1x_2$$

$$(3) \sum x_2y = \beta_0 \sum x_2 + \beta_1 \sum x_1x_2 + \beta_2 \sum x_2^2$$

Example 8.39: Recall the example 8.27 where the manager of a company tried to relate the productivity index with the intelligence score of the workers. Now suppose he wants to add one more variable i.e. the years of experience as another independent variable. The data are presented in the following table. Fit a multiple regression.

Model Aptitude score (x_1)	Productivity (y)	Years of experience (x_2)
78	79	2
65	60	8
73	62	5
48	40	6
53	52	3
70	80	5
72	85	4
65	62	6
60	68	4
61	60	8

Solution:

Let the productivity index, the dependent variable be expressed as y and the two independent or explanatory variables aptitude score and years of experience be x_1 and x_2 respectively.

The regression equation will be:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Three normal equations have to be solved here to get the values of β_0 , β_1 , β_2 .

Let us calculate the various values required for the normal equations.

y	x_1	x_2	$x_1 y$	$x_2 y$	$x_1 x_2$	y^2	x_1^2	x_2^2
79	78	2	6162	158	156	6241	6084	4
60	65	8	3900	480	520	3600	4225	64
62	73	5	4526	310	365	3844	5329	25
40	48	6	1920	240	288	1600	2304	36
52	53	3	2756	156	159	2704	2809	9
80	70	5	5600	400	350	6400	4900	25
85	72	4	6120	340	288	7225	5184	16
62	65	6	4030	372	390	3844	4225	36
68	60	4	4080	272	240	4624	3600	16
60	61	8	3660	480	488	3600	3721	64
648	645	51	42754	3208	3244	43682	42381	295

Normal Equations:

$$648 = 10\beta_0 + 645\beta_1 + 51\beta_2$$

$$42754 = 645\beta_0 + 42381\beta_1 + 3244\beta_2$$

$$3208 = 51\beta_0 + 3244\beta_1 + 295\beta_2$$

Solving these three equations we will get the estimated values of the coefficients as:

$$\beta_0 = -3.35$$

$$\beta_1 = 1.16$$

$$\beta_2 = -1.27$$

Thus, the estimated regression line will be

$$y = -3.35 + 1.16 x_1 - 1.27x_2$$

Interpretation

$$y = -3.35 + 1.16 x_1 - 1.27x_2$$

y = productivity Index

x_1 = Aptitude Score

x_2 = Years of Experience

From the regression equation it can be said that 1 unit increase in aptitude score will increase the productivity index by 1.16 unit. However, productivity will reduce by 1.127 units with the one unit increment of year of experience.

8.2.2.2 Regression with Dummy Variable

A dummy variable is a variable that takes the value either zero or one. The following example illustrates its application.

Example 8.40: The following table shows the gender wise monthly salary of teachers of a management school. Find out whether there is any gender bias in the salary structure of the school.

Salary (000 Rs.)

Female	Male
17	20.5
17.5	21.0
18	21.2
18.5	21.7
19.0	22.0

Let the regression equation be $Y_i = a + bD_i + e_i$

Where Y_i = the monthly salary of the teacher

D_i = the dummy variable showing the gender

$D_i = 1$ if male

$D_i = 0$ if female

Solution:

The data is coded and arranged as follows:

Salary (Y_i)	Male/Female (D_i)	$D_i Y_i$	D_i^2	$(D_i - \bar{D})^2$	Y_i^2
17.0	0	0	0	0.25	289
17.5	0	0	0	0.25	306.25
18.0	0	0	0	0.25	324
18.5	0	0	0	0.25	342.25
19.0	0	0	0	0.25	361
20.5	1	20.5	1	0.25	420.25
21.0	1	21.0	1	0.25	441
21.2	1	21.2	1	0.25	449.44
21.7	1	21.7	1	0.25	470.89
22.0	1	22.0	1	0.25	484
196.4	5	106.4	5	2.5	3888.08

From the above Table:

$$\Sigma Y_i = 196.4, \quad \Sigma D_i = 5, \quad \Sigma D_i Y_i = 106.4,$$

$$\Sigma D_i^2 = 5, \quad \Sigma (D_i - \bar{D})^2 = 2.5, \quad \Sigma Y_i^2 = 3888.08$$

The constants a and b are:

$$\begin{aligned} b &= \frac{n \Sigma D_i Y_i - \Sigma D_i \Sigma Y_i}{n \Sigma D_i^2 - (\Sigma D_i)^2} \\ &= \frac{1064 - 982}{50 - 25} \\ &= \frac{82}{25} \\ &= 3.28 \end{aligned}$$

Now, $a = \bar{y} - \hat{b}\bar{D}$

$$\bar{y} = \frac{196.4}{10} = 19.64$$

$$\bar{D} = 0.5$$

Thus,

$$\begin{aligned} a &= 19.64 - (3.28)(0.5) \\ &= 18 \end{aligned}$$

The estimate regression equation is

$$\hat{y}_i = 18 + 3.28D_i$$

From this result it can be said that the male teachers are getting 3.28 unit more salary than the female teachers. We now test for significance of gender bias in the salary structure.

The hypothesis is formulated as follows:

$$H_0: \beta = 0 \text{ (no gender bias)}$$

$$H_1: \beta \neq 0 \text{ (gender bias)}$$

The Test Statistic

$$\begin{aligned} t &= (b - \beta_0) \frac{\sqrt{(n-2) \Sigma(D_i - \bar{D})^2}}{\sqrt{\Sigma Y_i^2 - a \Sigma Y_i - b \Sigma D_i Y_i}} \\ &= \frac{3.28 \sqrt{8 \times 2.5}}{\sqrt{3888.08 - 3535.2 - 348.992}} \\ &= \frac{14.6686}{\sqrt{3.888}} \\ &= \frac{14.6686}{1.9728} \\ &= 7.44 \end{aligned}$$

Thus, $t = 7.44$.

Now, table value of t at 5% level of significance with $n - 2$ i.e. 8 degrees of freedom is 2.31. It indicates that gender bias is statistically significant.

Example 8.41: Now, if the years of experience is added to the example 8.40, what will happen? The data of years of experience is as follows:

Female	1	1	2	2	3
Male	2	3	4	3	5

Solution:

The regression equation will be

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 x_1 + e_i$$

where x_1 is the new independent variable years of experience.

The regression equation for this problem will be:

$$Y_i = 16.96 + 2.35D_i + 0.58x_1$$

The calculated t values for testing the significance of β_0 , β_1 and β_2 are:

For β_0 , $t = 57.72$

For β_1 , $t = 6.76$

For β_2 , $t = 3.97$

Tabulated value of t at 5% level of significance with $n - 3$ degree of freedom is 2.36. All the three are statistically significant at 5% level of significance. $\beta_2 = 0.58$, for example indicates for each additional year increase in experience, the average salary is predicted to increase by Rs.580.

8.3 CASELET

A little trading can push up m - cap in thinly traded scrips

Companies	M - Cap Change	Traded Value	XIER
Vijay Spinning	846	5	172
United	8059	86	94
TM Utilities	6984	156	45
Hind Corp	298	7	44
Sterile Ind	13760	325	42
Raj Ind	548	14	40
Arora Properties	1969	55	36
Lakshmi Cement	468	19	24

How much does it take to increase the market cap of a company? Sometimes, not much, it appears. An ETIG analysis shows that in thinly held and thinly traded scripts the market capitalization can go up by as much 10 - 50 times of traded value.

In other words, by investing very little money, the wealth of shareholders can go up immensely. This has implications for efficient functioning of stock markets. The ET intelligence Group tracked trading volumes of companies, which saw a sharp rise in share price from April, or about 27 trading days till middle May.

Vijay spinning, a company with FY06 net sales of Rs.50 crore, saw its market capital rise from Rs.70 crore to Rs.1545 crore on the back of the total trading of Rs.5 crore. So wealth for all shareholders put together increased 172 times the money invested. As per declared data, promoter holding is 69% in this company.

United, reportedly India's second largest builder, saw its market cap rise by Rs.8059 crore with a total trading of Rs.86 crore. The promoter holding here is 60%. The net worth of the promoters thus increased by Rs.4835 crore or over \$1 billion on this trading volume.

Similar examples abound. Sterile Industries saw a market cap increase of \$3 billion in 27 trading sessions with a trading volume of around \$70 million. The promoter holding in Sterile Industries is 78.5%. Promoter wealth went up by over Rs.10,000 crore or \$2.4 billion. (Adapted from Economic Times, May 17, 2006)

- (i) Establish a suitable regression equation for the future predict.
- (ii) Use regression techniques to analyze the present situation.

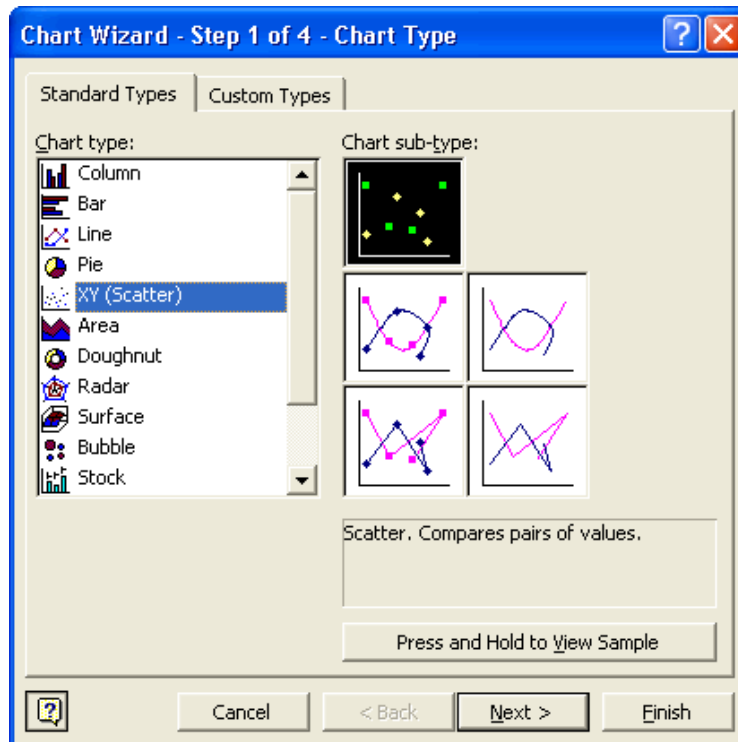
8.4 EXCEL GUIDE

Drawing of Scatter Diagram

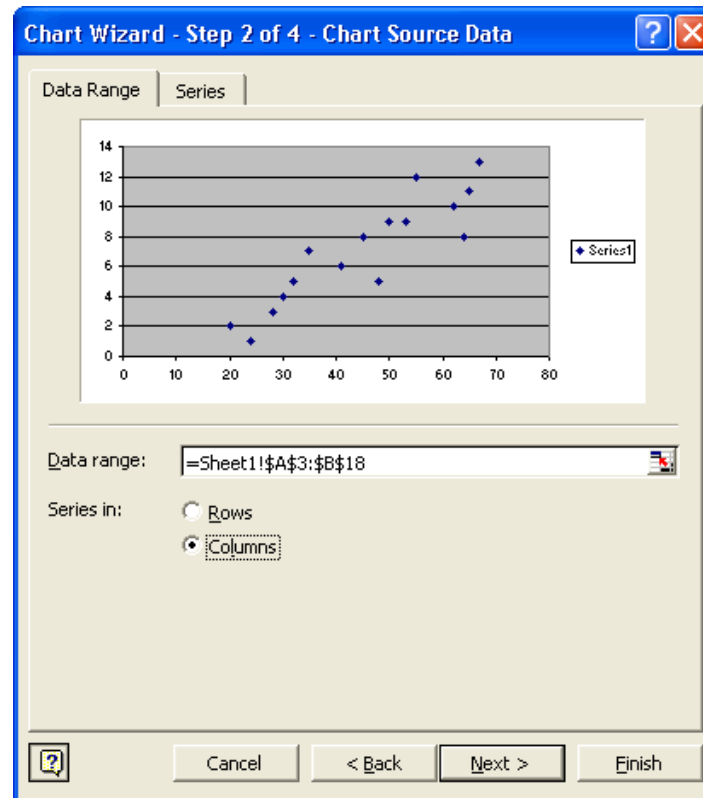
1. First, type the data in two columns. Age is in cells A3:A18 and Number of absences is in cells B3:B18.

Age	No. of absences
20	2
24	1
28	3
30	4
32	5
35	7
41	6
45	8
48	5
50	9
53	9
55	12
62	10
64	8
65	11
67	13

2. Select Insert > Chart from the tool bar to bring up the Chart Wizard.
3. Choose XY (Scatter) and select the unconnected points from the Chart sub-type



4. Click **Next>**
5. In the **Data Range** box, indicate where all of your data (the X and Y variables) are located. For example, you might indicate A3:B18. Since your data is in columns, check **Columns** under: “**Series in.**”



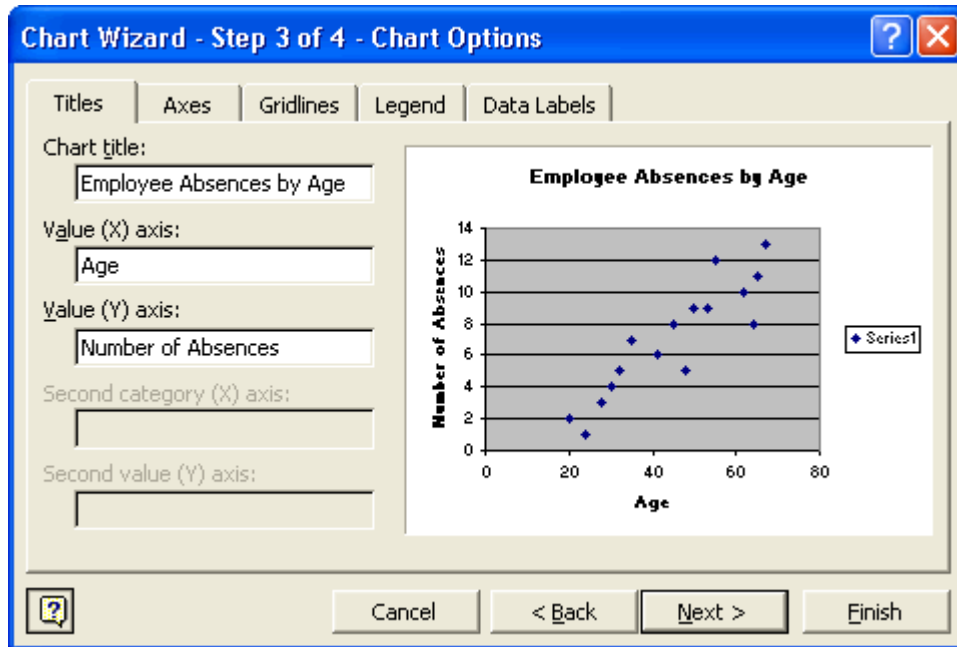
6. Click **Next>**

7. To now write labels for the chart:

Chart title: Give the chart a name, e.g., Employee Absences by Age

Value (X) Axis: variable name for the x-variable, e.g., Age

Value (Y) Axis: variable name for the y-variable, e.g., Number of Absences



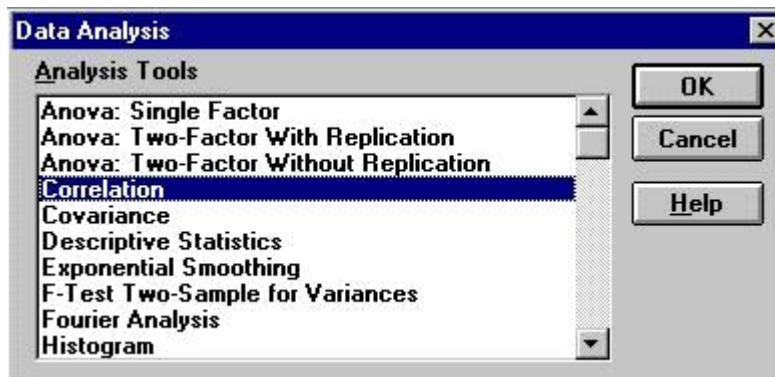
Calculation of Correlation Coefficient

x	y
1.0	2.6
2.3	2.8
3.1	3.1
4.8	4.7
5.6	5.1
6.3	5.3

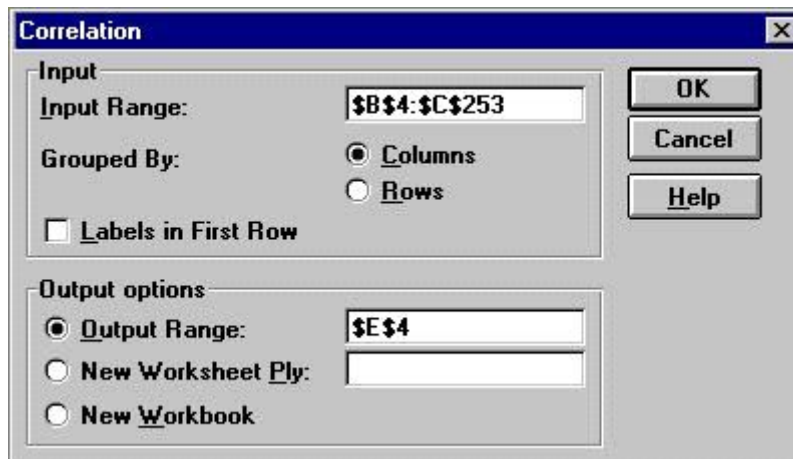
1. Open data file in Excel.
2. From the Tools menu, choose Data Analysis.



3. Scroll down until you see Correlation; highlight it, then click on OK.



4. Using the mouse, highlight the cells containing the data for both of the samples. Click on the circle labeled columns, assuming the data is arranged in that manner.



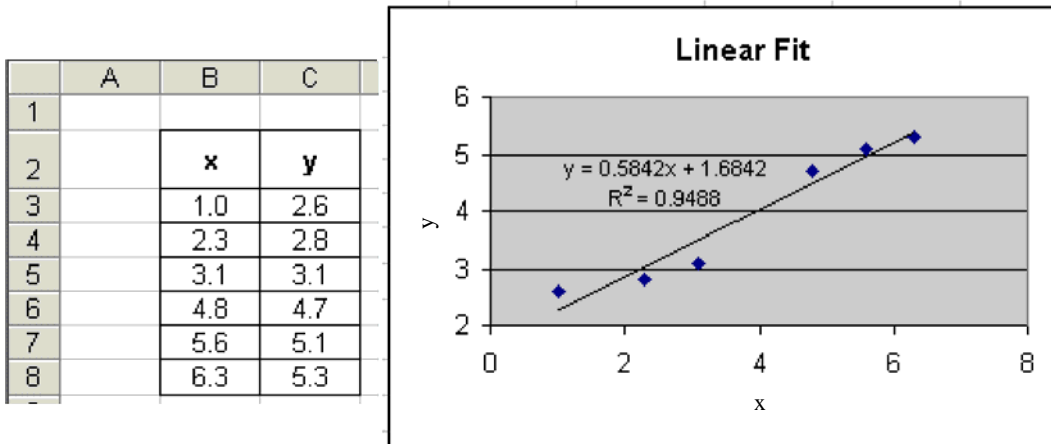
5. In the section labeled "Output Options", click on the circle beside output range; then click in the box beside it. Using the mouse, select a cell on your worksheet into which you would like the results to be placed, then click OK.

The resulting information will determine the correlation coefficient for the two data sets. The second box on the second row will give this coefficient.

Regression

Let's enter some data into an Excel spreadsheet, plot the data, create a trendline and display its slope, y-intercept and R-squared value. Recall that the R-squared value is the square of the correlation coefficient. (Most statistical texts show the correlation coefficient as " r ", but Excel shows the coefficient as " R ". Whether we write it as r or R , the correlation coefficient gives us a measure of the reliability of the linear relationship between the x and y values. (Values close to 1 indicate excellent linear reliability.)

1. Enter the data, suppose in columns B and C. The reason for this is strictly cosmetic as we will soon see.



Linear Regression Equations.

2. Given a set of data (x_i, y_i) with n data points, the slope, y-intercept and correlation coefficient, r , can be determined using the following:

$$b = \frac{n \sum(xy) - \sum x \sum y}{n \sum(x^2) - (\sum x)^2}$$

$$a = \frac{\sum y - m \sum x}{n}$$

$$r = \frac{n \sum(xy) - \sum x \sum y}{\sqrt{[n \sum(x^2) - (\sum x)^2][n \sum(y^2) - (\sum y)^2]}}$$

These quantities can be calculated as follows:

- (i) To count the number of data points n , the syntax is:

$$= \text{COUNT}(B3 : B8)$$

(ii) To calculate Σx , the syntax is:

$$= \text{SUM}(\text{B3}:\text{B8})$$

(iii) The syntax for xy column

$$= \text{PRODUCT}(\text{B3}:\text{C3})$$

and then use the AUTOFILL command

(iv) The syntax for x^2 and y^2 columns:

$$= \text{B3} \wedge 2 \text{ and}$$

$$= \text{C3} \wedge 2 \text{ respectively}$$

and then the AUTOFILL command.

(v) The summation terms can be calculated by using the SUM syntax.

(vi) Syntax for calculating the slope b , the y -intercept a and the correlation coefficient are shown in the following EXCEL Worksheet.

A11		=COUNT(B3:B8)									
	A	B	C	D	E	F	G	H	I	J	
1											
2		x	y	xy	x²	y²					
3		1.0	2.6	2.6	1.0	6.8					
4		2.3	2.8	6.44	5.3	7.8					
5		3.1	3.1	9.61	9.6	9.6					
6		4.8	4.7	22.56	23.0	22.09					
7		5.6	5.1	28.56	31.4	26.0					
8		6.3	5.3	33.39	39.7	28.1					
9											
10	n	Σx	Σy	$\Sigma (xy)$	$\Sigma (x^2)$	$\Sigma (y^2)$					
11	6	23.1	23.6	103.16	110.0	100.4					
12											
13		$(\Sigma x)^2$	$(\Sigma y)^2$								
14		533.61	556.96								
15											
16	slope, m =	0.5842					=(A11*D11-B11*C11)/(A11*E11-B14)				
17	y-int, b =	1.6842					=(C11-C16*B11)/A11				
18	r =	0.9741					=(A11*D11-B11*C11)/SQRT((A11*E11-B14)*(A11*F11-C14))				

8.5 EXERCISES

- 8.1 What do you mean by the term correlation? Clearly explain with suitable illustration.
- 8.2 Define correlation coefficient and point the properties of it.
- 8.3 Briefly explain the different types of correlation with their uses.
- 8.4 What do you mean by regression? Point out the usefulness of regression analysis in business problems.

- 8.5 Briefly explain the properties of regression coefficient and show that regression coefficient is independent of both change in origin and scale.
- 8.6 Distinguish correlation and regression. How to calculate regression coefficient with the help of correlation coefficient?
- 8.7 The figures in the following tables shows the rainfall(in inches) and wheat production(00 kg) of few regions of Punjab for 2005. Calculate Karl Pearson's Coefficient of correlation between rainfall and wheat production for the state.

Rainfall (Inches)	Wheat Production (00 kg)
21	20
23	1825
24	1926
20	27
19	28
24	32
22	25
19	22
21	31
20	30

- 8.8 Calculate coefficient of correlation form the following data

Length of service (year)	Annual income (000 Rs)
6	14
8	17
9	15
10	18
11	19
12	22
14	26
16	27
18	30
20	33

- 8.9. Find out the partial correlation between trial anxiety and the numbers of doctors visits controlling for family medical history from the following correlation matrix.

	Trail anxiety	Medical history	Doctors visits
Trial anxiety	1		
Medical history	0.20	1	
Doctors visits	0.35	0.15	1

- 8.10 In the following table, the recorded data are showing the test scores made by salesmen on an intelligence test and their weekly sales:

Salesmen	1	2	3	4	5	6	7	8	9	10
Test Score	40	70	50	60	80	50	90	40	60	60
Sales (000 Rs)	2.5	6	4	5	4	2.5	5.5	3	4.5	3

Calculate the regression equation of sale on test score and estimate the probable weekly sales volume if a salesman make a score of 100.

- 8.11 The cost of maintenance of a generator seems to increase with the years of use. The following data has been collected. Express this relation with a suitable linear model.

Year of use	Monthly cost (Rs)
5	900
4.5	840
4.5	821
4	600
4	723
3	547
3.5	580
4	835
1	475
2	700
1	462

- 8.12 Data on the annual sales of a company in lakhs of Rupees over the past twelve years is shown in the following table. Determine a suitable straight-line regression model and predict about the sale of 2007.

Year	Annual sale (Lakhs of Rupees)	Year	Annual sale (Lakhs of Rupees)
1994	1	2000	9
1995	5	2001	13
1996	4	2002	14
1997	7	2003	13
1998	10	2004	18
1999	8	2005	20

8.13 In a certain Examination 10 students obtained the following marks in Business Statistics and POM. Find Spearman's rank correlation coefficient.

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks in Business Statistics	70	60	82	48	32	65	40	88	73	64
Marks in POM	85	42	75	68	45	63	60	90	62	60

8.14 In the following table the recorded data showing the test scores made by salesmen on an intelligence test and their weekly sales are:

Salesmen	1	2	3	4	5	6	7	8	9	10
Test Score	55	67	55	60	88	50	85	40	60	60
Sales (000 Rs)	3	6	4	5	4	3.5	5.5	3	4.5	3

Calculate the regression equation of sale on test score and estimate the probable weekly sales volume if a salesman make a score of 90.

8.15 The large multi national company wants to study the relationship between sales and advertising expenditure. A considerable amount of the advertising budget is spent on television commercials and the balance goes to print media advertising. The following data is from eight randomly selected sales periods:

Y	X ₁	X ₂
180	3	5
220	4	10
150	2	8
230	5	12
209	4	11
186	3	10
250	5	12
172	2	8

Where Y = sales in million of rupees

X_1 = Magazine advertising in million of rupees

X_2 = Television commercials in million of rupees.

- Compute the regression coefficients.
- What is the significance of the two regression coefficients relative to the problem?
- Predict total sales for a period in which joint media and television expenditures are Rs.4 million & Rs.15 million respectively.

8.16 A company is the national distributor of Italian olive oil. They have already carried out a simple regression analysis on sales against advertising expenditures in the given sales region. The company decided to add another independent variable, namely, the population in sales region in order to account for variation in advertising expenditures. Variable Y is the sales in millions of rupees, variable x_1 represents the advertising in thousands of rupees and variable x_2 represents the population of the region in millions. Six sales regions were selected for the study and the data is presented in the following table.

Y	X_1	X_2
35	50	3.9
50	95	7.2
25	35	2.2
30	30	3.7
40	65	2.2
55	110	8.4

- Find the multiple regression equation.
- Find the standard error of the estimate.
- Find the unexplained variation for this data.
- Predict sales if Rs.1,00,000 is spent on advertising in a region with a population of six million.

8.17 A random study of some husbands in Mumbai was conducted by a social researcher from TISS who found that the family expenditure on food per year was a function of the number of persons in the family and the annual income of the family. The study resulted in the following relationship.

$$Y_c = 400 + 0.025 X_1 + 275 X_2$$

Where,

X_1 is the family income in rupees.

X_2 is the number of family members.

- Predict the food expenditure of a family of four with an annual income of Rs.3,20,000.
- Because of a promotion, the family income increases to Rs.4,00,000. How would it affect the food expenditure?

- 8.18 A researcher was interested in determining if there is any correlation between the creativity of the children and the creativity of their parents. Eight children and their parents were interviewed and ranked according to their creativity. The results are shown as follows:

Child	Parent
7	8
5	5
4	6
6	4
8	9
9	7
7	9
6	5

Calculate the degrees of association between the creativity ranking of the children and their parents.

- 8.19 The director of a large chain of home furnishing products would like to be able to predict the sales performance of sales employees based upon their sales experience. A sample of 12 salespersons are randomly selected and their annual sales figure (in lakhs of rupees) and their year of sales experience are recorded in the following table:

Salesperson	Experience in Years	Annual sales
1	3	25
2	3	15
3	2	10
4	2	12
5	5	75
6	5	65
7	4	30
8	3	90
9	3	40
10	4	45
11	8	100
12	11	65

- (a) State the regression equation and interpret the meaning of b_0 and b_1 .
- (b) Predict the annual sales of the salesperson who has 5 years of sales experience.
- (c) Would it be appropriate to predict the annual sales of an employee with 15 years of experience? Explain.
- (d) Calculate the coefficient of determination. Also give its interpretation.

8.20 A random survey of ten students was undertaken to determine the relationship between their scores in high school and their first year scores in college. The data has been summarized in the following figures:

$$\sum x = 31.4; \sum y = 27.6$$

$$\sum x^2 = 99.88; \sum y^2 = 78.96; \sum xy = 88.23$$

- (a) Calculate the regression coefficients b_0 and b_1 .
 - (b) For a student with a score of 2.8 in high school, what value of his score do you expect in the first year of college?
 - (c) Compute the standard error of the estimate.
 - (d) Compute the coefficient of correlation.
- 8.21 In economics, the demand function for a product is often estimated by the price charged for such a product. The quantity of new chocolates sold and the corresponding price charged at 10 stores for a one-week period is shown in the following table:

Store	Quantity	Price (Rs.)
1	225	25
2	250	25
3	280	20
4	290	20
5	310	15
6	340	15
7	350	15
8	350	10
9	360	10
10	380	10

- (a) Calculate the regression coefficients b_0 and b_1 .
- (b) How does the quantity sold change with lowering of price by one rupee each?
- (c) Predict the quantity expected to be sold if the price of the chocolate is fixed at Rs.10.

- (d) Calculate the explained and total variation.
- (e) Calculate the coefficient of determination and define the relationship between price and quantity.

8.22 You are given below the following information about advertisement and sales:

	Adv. Exp. (X) (Rs. Crores)	Sales (Y) (Rs. Crores)
Mean	20	120
Standard Deviation	5	25

Correlation coefficient +0.8.

- (i) Calculate the two regression equations.
- (ii) Find the likely sales when advertisement expenditure is Rs.25 crores.
- (iii) What would be the advertisement budget if the company wants to attain sales target of Rs.150 crores. **(MBA, DU, 1999)**
- 8.23 An industrial engineer collected the following data on experience & performance rating of 8 operators:

Operators	1	2	3	4	5	6	7	8
Experience (years)	16	12	18	4	3	10	5	12
Performance Rating	87	88	89	68	58	80	70	85

- (i) Does the data give evidence that experience improves performance?
- (ii) Estimate the performance rating of an operator having (a) 9 years and (b) 15 years of experience. **(MBA, MD Univ., 1994; MBA, Kumaun Univ., 2002)**
- 8.24 Regression calculations were carried out as follows:

$$\sum X = 32, \sum Y = 24, \sum XY = 218$$

$$\sum X^2 = 296, \sum Y^2 = 162.5, n = 4$$

Find the lines of regression and coefficient of correlation and comment

(MBA, MD Univ., 2000)

- 8.25 Compute the coefficient of correlation between the annual income (in thousand of rupees) and the amount of life insurance (in thousands of rupees) of eight families of the same size:

Annual Income	10	13	15	18	21	24	27	30
Amount of Insurance	15	12	25	20	25	30	35	32

8.26 In a research problem where two variables x and y were measured each 40 times, the following data were obtained:

$$\sum_{i=1}^{40} x_i = 293, \quad \sum_{i=1}^{40} y_i = 357, \quad \sum_{i=1}^{40} x_i^2 = 2685, \quad \sum_{i=1}^{40} x_i y_i = 3667$$

- (i) Find the slope and the y – intercept of the line of best fit.
- (ii) Give the least square regression line.

8.27 The scores of the final tests in marketing, HR and QT for eight randomly picked students are given below:

Marketing	50	58	67	70	75	82	86	92
HR	62	54	63	78	81	78	88	90
QT	79	82	70	74	67	62	64	56

- (i) Find the regression of scores in HR on scores in marketing.
- (ii) Find the regression of scores in QT on scores in marketing.
- (iii) Plot the scatter diagram for the HR – marketing scores and draw the fitted regression line.
- (iv) Do the same as in (iii) for the QT – marketing scores.



9

Time Series and Forecasting



Structure

- 9.1 Introduction
- 9.2 Goal of Time Series Analysis
- 9.3 Components of Time Series
 - 9.3.1 Secular Trend
 - 9.3.2 Seasonal Component
 - 9.3.3 Cyclical Component
 - 9.3.4 Random/Irregular Component
 - 9.3.5 Models of Time Series
- 9.4 Measurement of Secular Trend
 - 9.4.1 Free Hand Curve Fitting
 - 9.4.2 Semi Average Method
 - 9.4.3 Moving Average Method
 - 9.4.4 Fitting of a Straight Line/Trend Line
 - 9.4.5 Fitting of Exponential Trend
 - 9.4.6 Fitting of Second Degree Polynomial Equation
- 9.5 Measurement of Seasonal Component
 - 9.5.1 Method of Simple Average
 - 9.5.2 Ratio-to-Trend Method/Percentage-to-Trend-Method
 - 9.5.3 Ratio-to-Moving Average Method
- 9.6 Measurement of Cyclical Component
 - 9.6.1 Residual Method
- 9.7 Measurement of Irregular Component
- 9.8 Business Forecasting: An application of Time Series Analysis
 - 9.8.1 The Exponential Smoothing Method
 - 9.8.2 Trend adjusted for Seasonal Index Method
- 9.9 Caselets
- 9.10 Excel Guide
- 9.11 Exercises

9.1 INTRODUCTION

A time series is a sequence of observations, which are arranged in accordance of time. For example, hotel occupancy rates over a period of time, number of PC's sold for the last 7 years, sales of washing machines over the last 5 months, rate of air pollution for the past 5 years, weekly share prices, employment figures and numerous other such cases. Thus, a time series data relates a variable with time. Time series analysis is the analysis of such data to understand the behaviour of the series over a period of time.

Consider the following data:

Table 9.1
Sales of Washing Machines

Year	2002	2003	2004	2005	2006
Sales of Washing Machine (in 000)	200	250	300	320	350

Now consider the following equation to represent this data,

$$Y = f(t)$$

Let Y = sales of washing machine

t = time (years)

Then,

$Y = f(t)$ is a time series

A time series is best displayed in a scatter plot. Time is plotted on the horizontal axis and Y along the vertical axis.

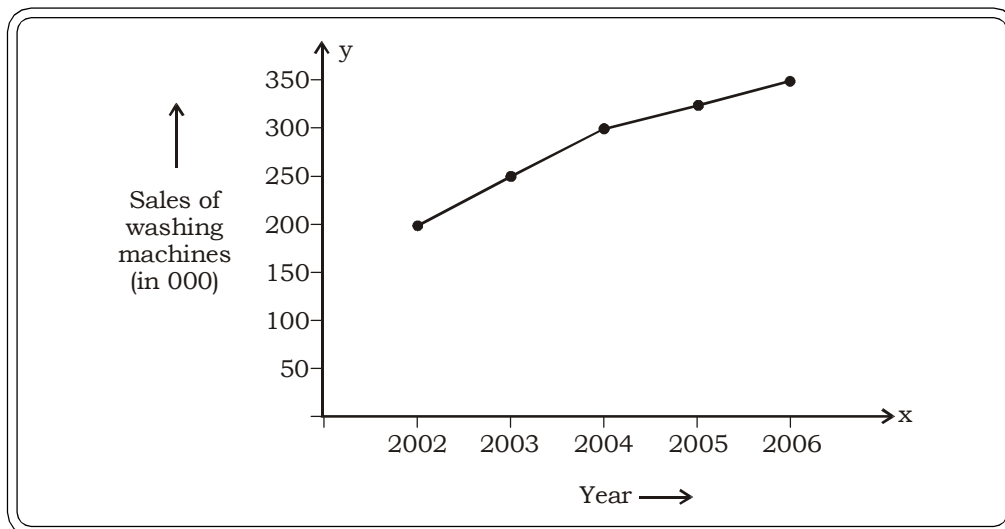


Fig. 9.1

A Time Series Plot

The above figure shows the time series plot of the data on sales of washing machines from 2002 to 2006. The graph seems to indicate a relationship between time and the sales of washing machines. The nature of the time series relationship is usually studied by decomposing the series into its component. These components are described in section 9.3

9.2 GOAL OF TIME SERIES ANALYSIS

There are basically two main goals of time series analysis viz.

- (i) to study the nature of the phenomenon represented by the sequence of observations and
- (ii) forecasting. Time series analysis is one of the most powerful and widely used forecasting methods. If a trend can be fitted and the rate of change can be ascertained, then one of the most important matters of the business world is to make estimates for the future which can easily be done by the time series analysis. Not only planning or estimation of future, time series also help to analyze the past in detail.

9.3 COMPONENTS OF TIME SERIES

There are different components of time series analysis depending on the kind of variability. Broadly it can be grouped into four categories:

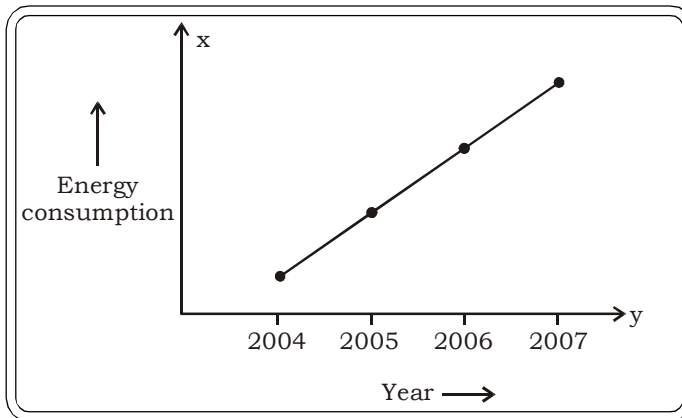
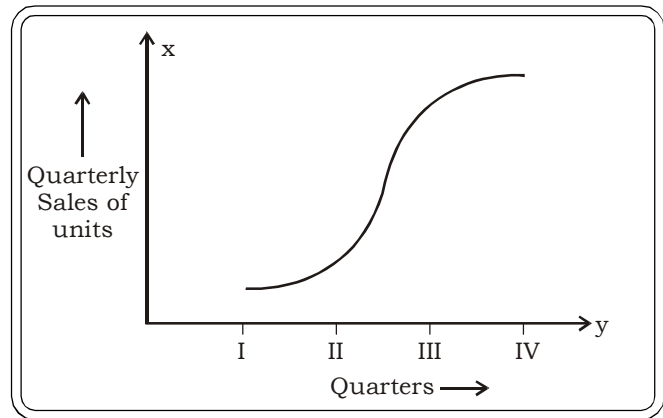
- Secular trend
- Cyclical Component
- Seasonal Component
- Random/Irregular component

9.3.1 Secular Trend

Normally, the tendency of the time series data to increase, decrease or stagnate over a long passage of time, is called trend or secular trend. Often, the trend component of a time series is a main feature and hence dominates the time series. Trend is usually, the result of long-term factors such as 5-year production, demographic characteristics of the population etc. It is thus the long-term movement in a time series. There are different types of secular trend. Broadly we can divide trend into two groups:

- Linear Trend
- Non-Linear Trend or Curvilinear Trend

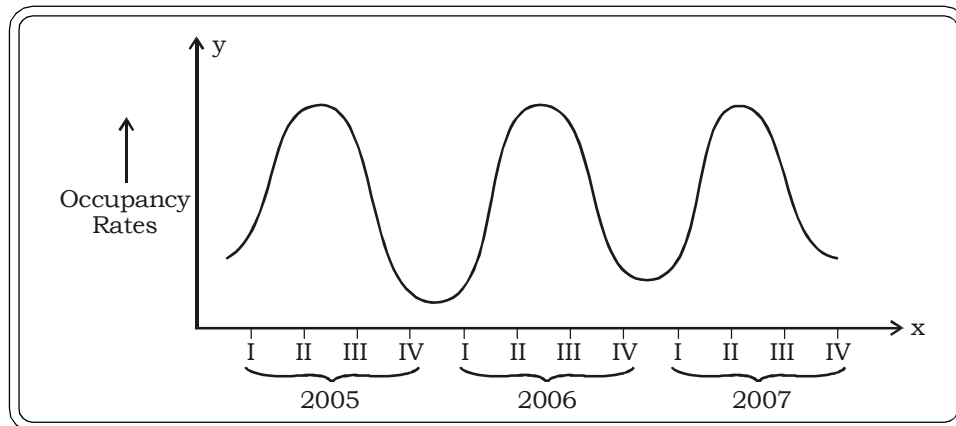
Figure 9.2 shows a linear trend between energy consumption and time and figure 9.3 shows a non-linear trend in the sales of a new product over time.

**Figure 9.2****Linear Trend****Figure 9.3****Non-Linear Trend**

9.3.2 Seasonal Component

Seasonal variation occurs because of the variability in the behavioral pattern during different seasons in a year. For example variability in sales of umbrellas is season dependent, sales of ice creams are more during the summer months and so on. However, the time intervals of occurrence of the seasonal variation are more or less uniform. The timing of the occurrence of seasonal variation has to be less than one year. Normally, seasonal variation occurs with certain interval of months. The seasonal fluctuation is quite prominent in the agricultural sector. The basic forces behind seasonal fluctuations are social customs, changing weather conditions etc.

Figure 9.4 shows seasonal variation in the occupancy rates of a hotel at a holiday resort.

**Figure 9.4****Seasonal Variation**

9.3.3 Cyclical Component

Cyclical component is almost synonymous with the business cycle. Actually business cycles reflect the upswing and downswing in the time series data that are sometimes observed over extended periods of time which could be several years. As such, business cycles are less predictable than the other components of the time series data. A business cycle typically consists of four phases each of which may sometimes last for years. The four phases are:

- (i) Recovery or Growth
- (ii) Prosperity
- (iii) Recession
- (iv) Depression

Figure 9.5 shows the four phases of the business cycle.

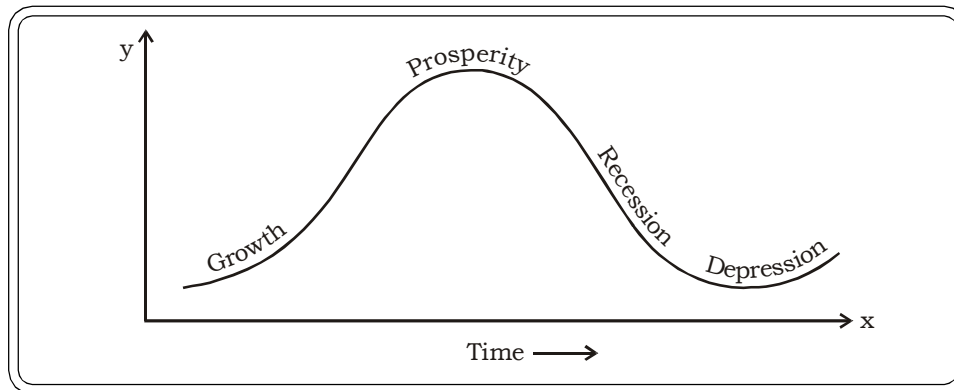


Figure 9.5

Business Cycle

9.3.4 Random /Irregular Component

This component consists of random variations that are very irregular in nature. This irregularity is mostly caused by random factors and sporadic causes such as strikes, floods etc. Alternatively, it is also the leftover component of the time series when the trend, seasonal and cyclical variation have been accounted for. Irregular variations usually occur over short intervals.

9.3.5 Models of Time Series

As time series data involves four components described in the earlier section, therefore analysis of time series essentially involves decomposition of the given data into the four components. The main objective here is to estimate and separate out the components.

Basically two popular models are there for time series decomposition viz.

- The Additive Model
- The Multiplicative Model

The Additive Model: In this model, the four components of the time series are considered to be additive in nature. It is mathematically represented as:

$$Y = T_t + C_t + S_t + R_t$$

where

Y = Time series

T_t = Secular trend

C_t = Cyclical Variation

S_t = Seasonal Variation

R_t = Random Variation

The additive model is used where it is assumed or expected that the components are independent of each other. Also, each component must have the same unit as the original data.

The Multiplicative Model: In this model, the components are in multiplicative form as given below

$$Y_t = T_t C_t S_t R_t$$

where the symbols have usual meaning.

It is assumed that the four factors are independent. Thus, the overall result is the combined effect of all the components.

This model is appropriate for situations when the amplitude of both seasonal and irregular variation increase as the level of trend rises. The multiplicative decomposition model can be transformed into the additive model by taking logarithms, which can be expressed as,

$$\text{Log } Y = \text{Log } T_t + \text{Log } C_t + \text{Log } S_t + \text{Log } R_t$$

9.4 MEASUREMENT OF SECULAR TREND

The long-term secular trend can be measured by fitting either linear trend or exponential or parabolic and so on. There are many methods of fitting the trend in time series. Some of these are:

- Free Hand Curve Fitting
- Method of Semi – Average
- Method of Moving Average
- Method of Least Square

9.4.1 Free Hand Curve Fitting

A free hand curve is fitted by inspection of the activity on the graph paper. Here the trend is basically a straight line drawn from the introspection of the analyst. Once the straight line is fitted, a trend equation $y = a + bx$ can easily be established in which

a is the intercept indicating the trend value when x is zero.

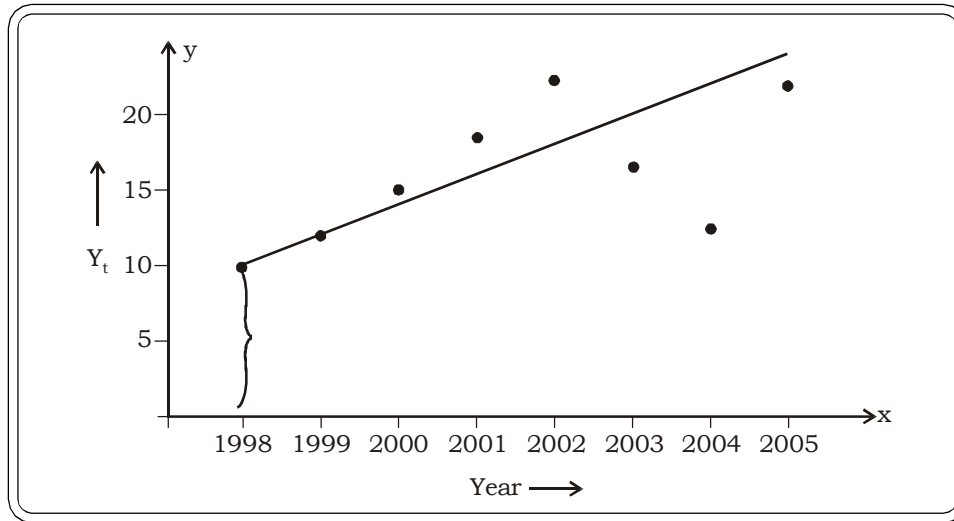
b is the slope of the trend line that indicates the change in time series variable due to one unit change in the independent variable.

The value of a can be obtained by measuring the distance on the vertical axis from where the straight line originates. Now the value of b is calculated by following two steps.

- (i) In the first step, the difference of the trend values at the first and last year is taken which actually gives the overall absolute change in the time series.
- (ii) Then dividing it by the number of years the value of b is calculated.

Example 9.1: Fit a free hand curve to the following data:

Year	1998	1999	2000	2001	2002	2003	2004	2005
Y_t	10	12	15	19	20	17	14	19

Solution:**Figure 9.6****Trend Line Fitted by Free Hand Method**

In figure 9.6, a trend line has been drawn approximately representing the given data. To fit a curve now of the form

$$y = a + bx$$

the value of a is the distance on the vertical axis from where the line originates. Thus

$$a = 10$$

Now, $b =$ difference of the trend values of the first and last year divided by the number of years.

$$\begin{aligned} &= \frac{19 - 10}{8} \\ &= 1.125 \end{aligned}$$

Thus, the trend line fitted by the free hand method is:

$$y = 10 + 1.125x$$

For forecasting a future value, the trend line drawn by this method is simply extended.

The main draw back of the free hand method is that it is highly subjective in nature. There are no fixed methods of drawing it. The line may vary from person to person. However, this method is very simple and gives an overall idea about the trend.

9.4.2 Semi Average Method

In this method, the total series of observations are subdivided into two parts. The average of each part is computed and placed against the middle period. If the data series is of odd order, to get the equal sub divided part the middle most items is selected. Now taking the two average points a curve is fitted, known as semi average curve. Further, assuming that, a linear function would adequately describe the data, a trend line may now be fitted. The estimation of the slope and the intercept is as follows:

- (i) The average of the first part is taken as the intercept i.e. assuming the line to be $Y = a + bx$.
- (ii) The slope is determined as follows:

$$b = \frac{\text{Difference in the arithmetic means of the two parts}}{\text{Difference in the time period between the means}}$$

Example 9.2: Fit a trend curve to the following data by the method of semi-averages and draw the semi-average curve:

Year	Annual Income (Lakhs, Rs.)
1980	1
1981	1.5
1982	1.9
1983	1.9
1984	1.95
1985	2.00
1986	2.10
1987	2.22
1988	2.31
1989	2.42
1990	2.5
1991	2.55
1992	2.10
1993	3.11
1994	3.25

Solution:

Since the data contains an odd number of years, the data is divided into two equal parts of 7 years each by ignoring the central value corresponding to the year 1987. The average of the first part centered at 1983 is 1.76 and the average of second part centered at 1991 is 2.72. By plotting these two points and joining them, we may obtain the trend line. This line can be extended for forecasting purposes. Now, to calculate the trend line

$$Y = a + bx$$

We consider $a = 1.76$ at 1983

$$\text{And } b = \frac{2.72 - 1.76}{1991 - 1983} = \frac{0.96}{8} = 0.12$$

Thus, the fitted trend line is

$$T_t = 1.76 + 0.12t$$

Using this line, forecasts can be made for future time periods, the underlying assumption being that the same trend is expected to continue upto that time period.

Year	Annual Income (Lakhs, Rs.)	Semi Average
1980	1	1.76
1981	1.5	
1982	1.9	
1983	1.9	
1984	1.95	
1985	2.00	2.72
1986	2.10	
1987	2.22	
1988	2.31	
1989	2.42	
1990	2.5	2.72
1991	2.55	
1992	2.10	
1993	3.11	
1994	3.25	

Figure 9.7 shows the semi-average curve

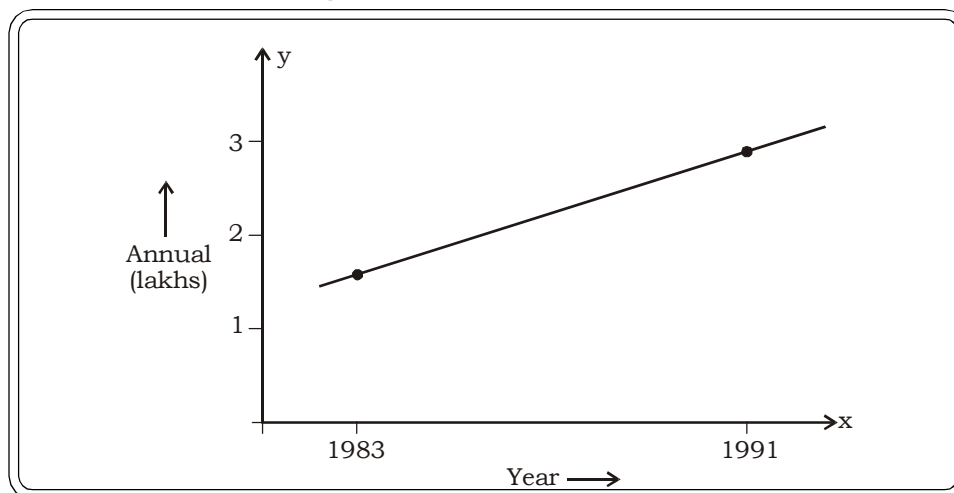


Figure 9.7

Semi - Average Curve

9.4.3 Moving Average Method

We have already defined a time series as a sequence of observations, which are ordered in time. While in long-term data, possibility of random fluctuation is almost certain, methods are also available to reduce the fluctuation for better prediction of the future. These methods of reducing random fluctuation of an observation are termed as smoothing. These methods, when applied properly, show more clearly and correctly the underlying trends. Some of the popular methods of smoothing are as follows:

- (i) Method of Moving Averages
- (ii) Method of Weighted Moving Averages.
- (iii) Method of Semi Averages.

The method of Semi Averages has been discussed in the previous section. In this section, the first two methods are discussed.

- (i) Method of Moving Averages

So far as ranking of the smoothing methods are concerned, moving average method is the most popular and widely used method of time series data processing. The underlined idea of this method is that the fluctuation or random means in the observation at the point of time has less impact on trend. This is practically implemented by taking the average of such points where fluctuation occurs, with the values of the series immediately preceding it and immediately following it.

A moving average can be obtained by successively averaging overlapping groups of two or more consecutive values in a time series.

Depending upon the number of points considered, moving average methods can be of different periods like: three yearly moving average, four yearly moving average, five yearly moving average and so on.

For example, consider the following data related to sales of ice creams (in thousands).

Table 9.2
Sales of Ice-Cream (in thousands)

Year	I	II	III	IV
2004	42	58	80	60
2005	46	60	82	64
2006	44	56	85	70
2007	48	54	89	72

The four year moving average computations are shown in the following table.

Table 9.3
Centered Four-Quarterly moving Averages

Year	Quarter	Sales	Four Quarterly Moving Averages	Centered Four Quarterly Moving Averages
2004	I	42	{ 60 } { 61 } 61.5	
	II	58		
	III	80		60.5
	IV	60		61.25
2005	I	46	62	61.75
	II	60	63	62.5
	III	82	62.5	62.75
	IV	64	61.5	62.00
2006	I	44	62.25	61.875
	II	56	63.75	63
	III	85	64.75	64.25
	IV	70	64.25	64.5
2007	I	48	65.25	64.75
	II	54	65.75	65.5
	III	89		
	IV	72		

In the above example, since the number of observations considered for each moving average does not coincide with an observed data, therefore, this necessitated the calculation of *centered moving averages* by averaging two four quarterly moving averages. These values are indicated in the last column of the above table and the method of obtaining them is explained below.

Centered moving averages (for even number of years). When an average of even order needs to be computed, the term centering becomes very important. Each average data should be placed at the center of the selected data set. However in this case, the moving average would not correspond to a particular time period and further analysis would become difficult. Therefore another set of centered moving averages is calculated by taking the average of two of the previously calculated moving averages and placing the value in between the two previous averages, as has been shown in the table above.

Centering is not necessary when the number of observations is odd.

It may be noted that there is a loss of information of 1st two quarters in 2004 and the last two quarters of 2007.

Example 9.3: From the following table showing 10-year data of sales of a company, find out the three yearly moving averages. Also compare the graphs.

Year	Sales (lakh Rs.)
1996	20
1997	30
1998	32
1999	50
2000	28
2001	22
2002	40
2003	41
2004	45
2005	47

Solution:

Year	Sales (lakh Rs.)	Three yearly total	Three yearly average
1996	20	-	-
1997	30	82	27.33
1998	32	112	37.33
1999	50	110	36.33
2000	28	100	33.33
2001	22	90	30
2002	40	103	34.33
2003	41	126	42
2004	45	133	44.33
2005	47	-	-

The graph in figure 9.8 shows the actual sales values and the smoothed values by the method of moving averages.

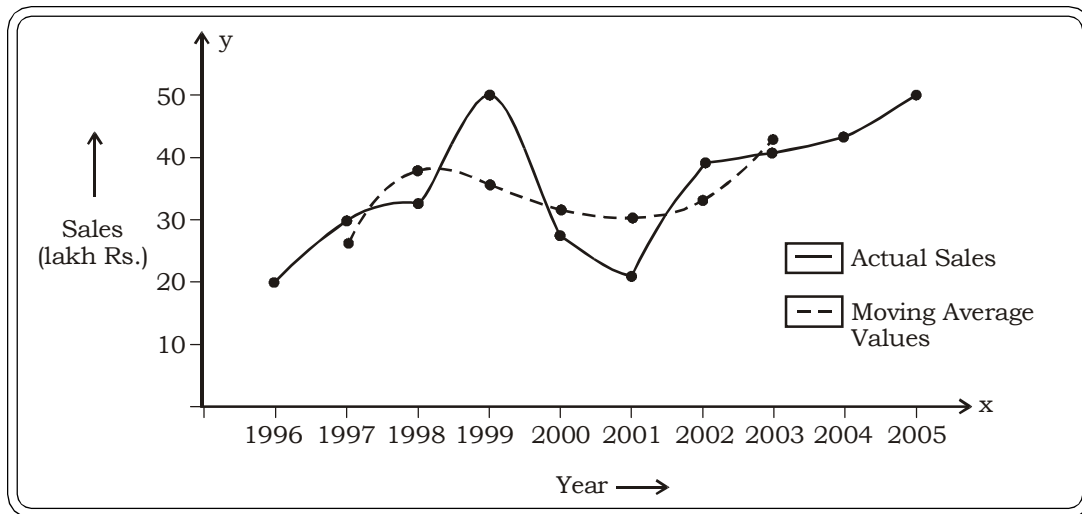


Figure 9.8

Comparison of Actual Values and Moving Average Values

Example 9.4: Calculate five yearly moving averages from the following data of number of cars sold yearly from 1992 to 2006.

Years	Sales (in 00,000)
1992	2
1993	1
1994	3
1995	2
1996	2
1997	2
1998	0
1999	4
2000	3
2001	2
2002	2
2003	1
2004	6
2005	2
2006	1

Solution:

Calculation of five yearly moving average

Years	Sales (in 00, 000)	Five yearly moving total	Five yearly moving average
1992	2		
1993	1		
1994	3	10	2
1995	2	10	2
1996	2	9	1.8
1997	2	10	2
1998	0	11	2.2
1999	4	11	2.2
2000	3	11	2.2
2001	2	12	2.4
2002	2	14	2.8
2003	1	13	2.6
2004	6	12	2.4
2005	2		
2006	1		

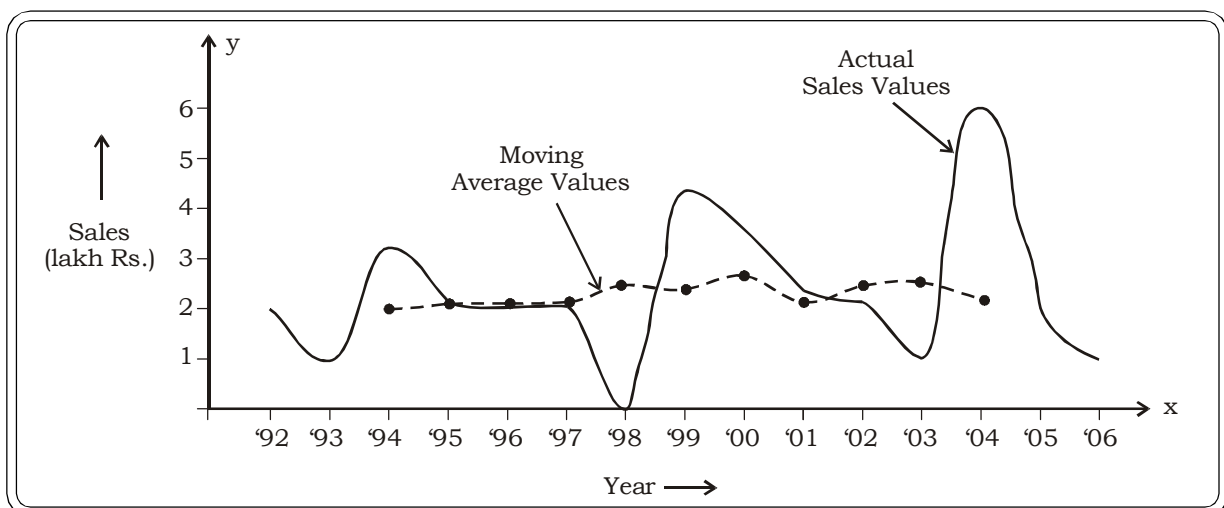


Figure 9.9

Comparison of Actual and Moving Average Values

Example 9.5: The following data is from a Travel Agency- Easy Trips, giving the gross revenues generated by the agency from 1997 to 2006 in millions of Rupees.

Years	Revenues
1997	4
1998	5
1999	7
2000	9
2001	12
2002	14
2003	17
2004	20
2005	21
2006	22

Calculated the four yearly moving averages.

Solution:

The centered moving average calculations are shown in the following table.

Year	Revenues (in millions Rs.)	Four yearly moving totals	Four yearly MA	Four Yearly centered MA
1997	4			
1998	5			
1999	7	25	6.25	7.25
2000	9	33	8.25	9.375
2001	12	42	10.5	11.75
2002	14	52	13	14.375
2003	17	63	15.75	16.875
2004	20	72	18	19
2005	21	80	20	
2006	22			

The first 4 yearly-centered moving average corresponding to 1999 is calculated as follows:

$$(6.25 + 8.25)/2$$

(ii) Method of Weighted Moving Averages

The moving average value assigns equal weightage to all values. However, it is also possible to assign different weights to different values as seen necessary. There are no fixed rules of assigning weights. One method is to assign more weightage to recent observations and less weightage to past observations, a principle similar to the Markovian principle.

The weighted moving average is calculated by using the following formula:

$$\text{Weighted Moving Average} = \sum \frac{(\text{weight for time period } t)(\text{observation for time period } t)}{\text{Sum of the weights}}$$

An example of weighted moving average is now discussed.

Example 9.6: The manager of a retail store wants to forecast sales of a particular brand of soap, which was recently introduced by a company. Sales data of the last 12 weeks is available to the manager.

Weeks	1	2	3	4	5	6	7	8	9	10	11	12
Sales (in 000)	2	3	4	3	4	6	8	10	12	11	13	14

The manager decides to assign the following weights to calculate the three-weekly weighted average.

Weeks	Weights
Last week	4
Two weeks ago	3
Three weeks ago	2
Total	9

Solution:

Since the weights are for three previous weeks, moving average values of the first three weeks cannot be computed. The three period weighted moving average values starting from the fourth week are as follows:

Weeks	Sales (in 000)	3-week weighted moving average
1	2	–
2	3	–
3	4	–
4	3	$\frac{4 \times 4 + 3 \times 3 + 2 \times 2}{9} = \frac{16 + 9 + 4}{9} = 3.22$
5	4	$\frac{3 \times 4 + 4 \times 3 + 3 \times 2}{9} = \frac{12 + 12 + 6}{9} = 3.33$
6	6	$\frac{4 \times 4 + 3 \times 3 + 4 \times 2}{9} = \frac{16 + 9 + 8}{9} = 3.67$
7	8	$\frac{6 \times 4 + 4 \times 3 + 3 \times 2}{9} = \frac{24 + 12 + 6}{9} = 4.67$
8	10	$\frac{8 \times 4 + 6 \times 3 + 4 \times 2}{9} = \frac{32 + 18 + 8}{9} = 6.44$
9	12	$\frac{10 \times 4 + 8 \times 3 + 6 \times 2}{9} = \frac{40 + 24 + 12}{9} = 8.44$
10	11	$\frac{12 \times 4 + 10 \times 3 + 8 \times 2}{9} = \frac{48 + 30 + 16}{9} = 10.44$
11	13	$\frac{11 \times 4 + 12 \times 3 + 10 \times 2}{9} = \frac{44 + 36 + 20}{9} = 11.11$
12	14	$\frac{13 \times 4 + 11 \times 3 + 12 \times 2}{9} = \frac{52 + 33 + 24}{9} = 12.11$

9.4.4 Fitting of a Straight Line/Trend Line

A straight-line trend is appropriate when the growth of a time series is relatively a constant amount. To fit the straight-line trend we will apply least square method. The method of least square is already discussed in chapter 5 where this method was used for fitting the regression equation. Here briefly we will discuss it. Let us assume the straight line be $y = a + bx$

where $x \rightarrow$ denotes the time variable in this case

$y \rightarrow$ The observations at different points of time

a and b are constants to be determined.

As we have two parameters a and b we will have two normal equations viz.

$$\sum y = na + b \sum x$$

$$\sum xy = n \sum x + b \sum x^2$$

By solving these two equations we can get the values of a and b , and thus obtain the trend line.

Examples 9.7: Below are given the figures related to profit of a business firm from 1999 to 2005.

Year	2000	2001	2002	2003	2004	2005	2006
Profit (00,000 Rs.)	30	35	40	42	45	48	50

- (i) Find a straight-line trend to this data.
- (ii) Make an estimate of profit for the year 2006.
- (iii) Also, draw the trend line.

Solution:

Year (x)	Profit Y	Deviations from Middle Year (X): $X = x - 2003$	XY	X^2	Trend Value (Approx)
2000	30	-3	-90	9	32
2001	35	-2	-70	4	35
2002	40	-1	-40	1	38
2003	42	0	0	0	41
2004	45	1	45	1	44
2005	48	2	96	4	48
2006	50	3	150	9	51
	$\sum y = 290$	$\sum X = 0$	$\sum XY = 91$	$\sum X^2 = 28$	

The equation of the straight line is

$Y = a + bX$, where Y denotes the profit and x denotes the year.

Now, by applying least square method we will get the two normal equations as

$$\sum Y = na + b \sum X$$

$$\sum xy = n \sum X + b \sum X^2$$

$$\sum X = 0; \sum X^2 = 28; \sum Y = 290$$

$$n = 7; \sum XY = 91$$

Thus, the first normal equation

$$\sum Y = na + b \sum X \text{ or}$$

$$\Rightarrow 290 = 7a$$

$$\Rightarrow a = \frac{290}{7} = 41.43$$

The second normal equation

$$\sum XY = a \sum X + b \sum X^2$$

$$\Rightarrow 91 = b \times 28$$

$$\Rightarrow b = \frac{91}{28} = 3.25$$

(i) Thus the fitted equation is

$$Y = 41.43 + 3.25 X$$

(ii) During 2006, $X = 4$

$$\begin{aligned} \text{Thus, } Y &= 41.43 + 3.25 \times 4 \\ &= 41.43 + 13 \\ &= 54.43 \end{aligned}$$

(iii) Figure 9.10 below shows the fitted trend line

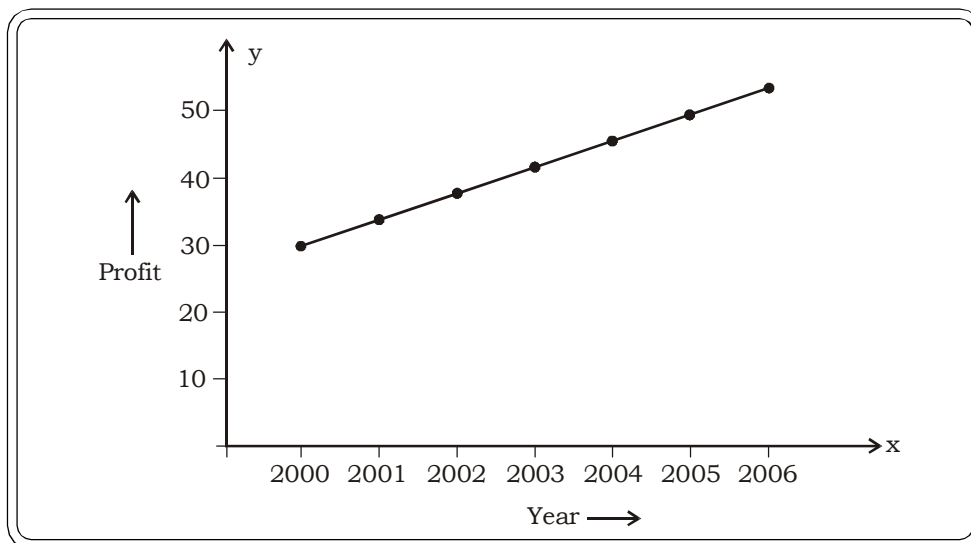


Fig. 9.10

Trend Line

Example 9.8: Fit a linear trend line to the following data:

Year	1985	1986	1987	1988	1989	1990
Sales (in lakhs)	15	25	30	40	42	45

Solution:

Year	Profit Y	Taking Deviation X	XY	X ²
1985	15	-5	-75	25
1986	25	-3	-75	9
1987	30	-1	-30	1
1988	40	1	40	1
1989	42	3	126	9
1990	45	5	225	25
	$\sum Y = 197$	$\sum X = 0$	$\sum XY = 211$	$\sum X^2 = 70$

Let the linear equation be

$$Y = a + bx$$

$$a = \frac{\sum Y}{n}; \text{ as } \sum X = 0$$

$$= \frac{197}{6} = 32.83$$

$$b = \frac{\sum XY}{\sum X^2}$$

$$= \frac{211}{70} = 3.01$$

The trend line fitted to the given data is

$$Y = 32.83 + 3.01 X$$

9.4.5 Fitting of Exponential Trend

Exponential trend is applicable where growth in time series data is nearly at a constant rate with respect to per unit time.

The exponential curve is given by the equation

$$Y_t = ab^X$$

Here a and b are the constants that need to be determined. Y and X are the dependent and independent variables respectively. This exponential equation can be transformed into a linear equation by taking logarithms on both sides as follows:

$$\log Y_t = \log a + X \log b$$

Here, the normal equations will be the following:

$$\sum \log(Y) = n(\log a) + (\log b) \sum X$$

$$\sum X(\log Y) = (\log a) (\sum X) + (\log b) (\sum X^2)$$

Now, if we set the point of origin at the middle of the given time series data, the parameter will be calculated as

$$\log a = \frac{\sum (\log Y)}{n}$$

$$\log b = \frac{\sum X(\log Y)}{\sum X^2}$$

Example 9.9: The index of industrial production in India during 10 years period from 1975 to 1985 are presented in the following table. Find exponential trend equation.

Year	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985
Index	100	115	130	137	135	130	140	148	155	162	180

Solution:

Year	Index Y	X	Log Y	X Log Y	X ²
1975	100	-5	2.21	-11.05	25
1976	115	-4	2.26	-9.02	16
1977	130	-3	2.11	-6.34	9
1978	137	-2	2.14	-4.27	4
1979	135	-1	2.13	-2.13	1
1980	130	0	2.11	0	0
1981	140	1	2.15	2.15	1
1982	148	2	2.17	4.34	4
1983	155	3	2.19	6.58	9
1984	162	4	2.21	8.83	16
1985	180	5	2.26	11.28	25
			23.93	0.36	110

The exponential curve is given by $Y = ab^X$

To calculate a & b, the normal equations are

$$23.93 = 11 \log a + 0. \log b$$

$$0.36 = 0. \log a + 110. \log b$$

$$\log a = \frac{\sum(\log Y)}{n} = \frac{23.93}{11} = 2.18$$

$$\log b = \frac{\sum X(\log Y)}{\sum X^2} = \frac{0.36}{110} = 0.0033$$

The exponential trend is

$$\text{Log } Y_t = 2.18 + 0.0033 X$$

9.4.6 Fitting of Second Degree Polynomial Equation

A second degree polynomial is used to model the time series when the scatter plot reveals curvature or the series changes direction once. The formula for second degree polynomial is

$$y_t = b_0 + b_1 t + b_2 t^2 + \varepsilon$$

Using the method of least squares, the normal equations for estimating the parameters b_0 , b_1 and b_2 are as follows:

$$\Sigma y = n b_0 + b_1 \Sigma t + b_2 \Sigma t^2$$

$$\Sigma t y = b_0 \Sigma t + b_1 \Sigma t^2 + b_2 \Sigma t^3$$

$$\Sigma t^2 y = b_0 \Sigma t^2 + b_1 \Sigma t^3 + b_2 \Sigma t^4$$

Solving these equations, we can estimate the values of b_0 , b_1 and b_2 and obtain the estimated second degree polynomial trend as

$$\hat{y}_t = \hat{b}_0 + \hat{b}_1 t + \hat{b}_2 t^2$$

These calculations may be simplified by coding the data or changing the scale such that

$$\Sigma t = 0 \quad \text{and} \quad \Sigma t^3 = 0$$

In such a case, the values of b_0 , b_1 and b_2 are given by the following formulae:

$$b_0 = \frac{\Sigma y - c \Sigma t^2}{n}; \quad b_1 = \frac{\Sigma t y}{\Sigma t^2} \quad \text{and} \quad b_2 = \frac{n \Sigma t^2 y - \Sigma t^2 \Sigma y}{n \Sigma t^4 - (\Sigma t^2)^2}$$

Example 9.10: Sugar production in (000 tonnes) during the years 1998 – 2004 are given below. Fit a non linear trend of the form $Y = a + bx + cx^2$ to the data

Year	1998	1999	2000	2001	2002	2003	2004
Sugar Production (0000 tonnes)	58	60	70	77	85	56	50

Obtain an estimate of sugar production in 2005.

Solution:

The equation to be fitted is $Y = a + bx + cx^2$

For calculating a, b and c, we use the following table:

Year	Sugar Production (y_t)	t Taking 2001 = 0	t^2	ty	t^4	t^2y
1998	58	-3	9	-174	81	522
1999	60	-2	4	-120	16	240
2000	70	-1	1	-70	1	70
2001	77	0	0	0	0	0
2002	85	1	1	85	1	85
2003	56	2	4	112	16	224
2004	50	3	9	150	81	450
	$\Sigma y_t = 456$	$\Sigma t = 0$	$\Sigma t^2 = 28$	$\Sigma ty = -17$	$\Sigma t^4 = 196$	$\Sigma t^2y = 1591$

From the table,

$$\Sigma y = 456, \Sigma t = 0, \Sigma t^2 = 28, \Sigma ty = -17, \Sigma t^4 = 196, \Sigma t^2y = 1591$$

Thus,

$$b = \frac{-17}{28} = -0.6071$$

$$c = \frac{7(1591) - (28)(456)}{7(196) - (28)^2}$$

$$= \frac{11137 - 12768}{1372 - 784} = \frac{-1631}{588} = -2.77$$

$$\begin{aligned}
 a &= \frac{\Sigma y - c\Sigma t^2}{n} \\
 &= \frac{456 - c \times 28}{7} \\
 &= \frac{456 + 77.56}{7} \\
 &= 76.22
 \end{aligned}$$

Thus the fitted parabolic equation:

$$y = 76.22 - 0.6071t - 2.77 t^2$$

For estimating sugar production of 2005, $t = 4$

$$\begin{aligned}
 y &= 76.22 - 0.6071 \times 4 - 2.77 \times 4^2 \\
 &= 29.47 \text{ (0000 tonnes)}
 \end{aligned}$$

9.5 MEASUREMENT OF SEASONAL COMPONENT

Seasonal variation as already defined, is that movement of a time series where the changes occur during a one year period or even less than a year, for example, over days, weeks or months. Detecting and measuring the seasonal variation of a time series data can be useful in many ways.

- (i) Firstly, it can help in analyzing the behaviour of the series in the past.
- (ii) Secondly, using this information, we can make projections for the future. Such projections could mean the ability to predict future patterns based on examination of past patterns.
- (iii) Thirdly, once the seasonal variation has been calculated, we can eliminate this from the time series and determine cyclical patterns in the data. This is known as deseasonalization of the data.

Some of the different methods of measuring the seasonal component of time series data are

- (i) Method of Simple Averages
- (ii) Ratio-to-Trend Method
- (iii) Ratio-to-Moving Average Method

9.5.1 Method of Simple Averages

The steps for calculating the seasonal index by this method are

- (i) The data is first arranged according to years, months or quarters as given.
- (ii) The totals for each month is computed
- (iii) Then, the average for each month is calculated. This is done by dividing the totals obtained in step (ii) by the number of months for which it has been computed.
- (iv) Next, we obtain the average of the monthly averages.

(v) The seasonal indices for each month are now calculated using the formula.

$$\text{Seasonal Index for Month } X = \frac{\text{Monthly Average for } X}{\text{Average of monthly averages}} \times 100$$

Where X = January, February, March,, December.

Example 9.11: Calculate the seasonal indices for the following data related to sales of cars (in thousands) in a city during 2004-2007. Use the method of simple averages.

Months	Years			
	2004	2005	2006	2007
January	364	394	399	347
February	342	367	379	325
March	288	345	328	302
April	262	309	360	270
May	236	284	300	247
June	245	279	308	230
July	249	251	270	220
August	268	269	310	230
September	250	287	300	250
October	300	320	340	279
November	328	328	350	280
December	299	367	390	310

Solution:

The monthly totals, the monthly averages and the seasonal indices are given in the following table.

Months	2004	2005	2006	2007	Monthly Totals	Monthly Averages	Seasonal Indices
January	364	394	399	347	1504	376	126.37
February	342	367	379	325	1413	353.25	118.72
March	288	345	328	302	1263	315.75	106.120
April	262	309	360	270	1201	300.25	100.83
May	236	284	300	247	1067	266.75	89.65
June	245	279	308	230	1026	265.5	89.23
July	249	251	270	220	990	247.5	83.16
August	268	269	310	230	1077	269.25	90.49
September	250	287	300	250	1087	271.75	91.33
October	300	320	340	279	1239	309.75	104.10
November	328	328	350	280	1286	321.5	108.05
December	299	367	390	310	1366	341.5	114.77

$$\text{Average of the monthly averages} = \frac{3570.45}{12} = 297.54$$

Example 9.12: The following data is related to the number of patients admitted to a hospital (in hundreds) during the period 2000-2004. Calculate the seasonal indices for each quarter.

Quarter	Years				
	2000	2001	2002	2003	2004
Q ₁	74	90	90	93	95
Q ₂	62	67	82	78	80
Q ₃	55	65	77	72	75
Q ₄	60	80	85	90	87

Solution:

The seasonal indices are computed as follows:

Quarters	2000	2001	2002	2003	2004	Quarterly Averages	Seasonal Indices
Q ₁	74	90	90	93	95	88.4	$\frac{88.4}{77.85} \times 100 = 113.55$
Q ₂	62	67	82	78	80	73.8	$\frac{73.8}{77.85} \times 100 = 94.41$
Q ₃	55	65	77	72	75	68.8	$\frac{68.8}{77.85} \times 100 = 88.38$
Q ₄	60	80	85	90	87	80.4	$\frac{80.4}{77.85} \times 100 = 103.27$
Average of Quarterly Averages						77.85	

The method of simple averages, though computationally easy, is not very useful. This is because, it assumes that the trend component is absent from the time series, when in reality this is rarely the case. Thus, the seasonal index calculated by this method actually reflects both trend and seasonality. The next method is considered to be better than the method of simple averages.

9.5.2 Ratio-to-Trend Method or Percentage-to-Trend Method

This method is an improvement over the method of simple averages. It assumes the presence of trend but eliminates it by isolating the seasonal factors as follows:

The basic multiplicative time series model $Y_t = T \times S \times C \times I$ reduces to

$$Y_t = \frac{T \times S \times C \times I}{T} = S \times C \times I$$

Thus, this method works very well, particularly if cyclical effects and random fluctuations are not very pronounced.

The model then further reduces to

$$\frac{Y_t}{C} = S \times I$$

Thus, the best results of this method can be obtained for data that are monthly or of even shorter intervals, thus providing a safety margin in assuming that cyclical components are absent.

The steps in the execution of the ratio-to-trend method are as follows:

- (i) The first step is to define the time unit to be used in the analysis. The time unit is generally monthly or quarterly or could be in other smaller units like hours and minutes also.
- (ii) The next step is to estimate the trend equation. The most common method is the usual least squares method.

- (iii) The trend equation is then used to estimate the trend values (T_t) for each time unit t .
- (iv) Next, we may calculate each observed value in the series (Y_t) for each time unit as a

percentage of the corresponding trend value i.e. $\left(\frac{Y_t}{T_t}\right) \times 100$

This step ensures that the trend values have been eliminated from the time series.

- (v) In the next step, we can determine whether or not there is a seasonal effect in the time series. For this, we need to examine the ratios $\left(\frac{Y_t}{T_t}\right) \times 100$. There are two indicators of the presence of seasonal effect.

(a) The period-by-period ratios $(Y_t/T_t) \times 100$'s are similar. For example, for monthly seasonal variation, the ratios of say July of one year should be similar to the ratio of July of other years.

(b) Secondly, the ratios $\left(\frac{Y_t}{T_t}\right) \times 100$ of one period should be different from the ratio of at least one of the other periods. For example the ratio for July should be different from the ratio $\left(\frac{Y_t}{T_t}\right) \times 100$ of at least one more month. If the ratios are same for all the months, then obviously the data does not display any seasonal effects.

- (vi) The next step involves calculating the median or modified mean of each period's $\left(\frac{Y_t}{T_t}\right) \times 100$.

For example, if the data is given in months for four consecutive years, then we calculate the median or modified mean for each month over the four years.

This step is executed to control irregular variations. Irregular movements produce extremes and since the median is robust to extreme values, its use is justified.

Alternatively, the highest and lowest values of a period are discarded, and a "modified mean" is calculated from the remaining values. This is done by assuming that irregular variations produced the highest and the lowest observations. The medians or modified means are the new seasonal indices.

- (vii) The final step involves adjusting the seasonal indices in such a manner that the average should be 100. This may be done by multiplying each index by a adjustment factor as follows:

$$\text{Adjustment Factor} = \frac{100}{\text{mean of the unadjusted indexes}}$$

Final Interpretation of the Seasonal Indices

A seasonal index exceeding 100 implies a positive seasonal effect and one below 100 implies a negative seasonal effect.

We now consider an example to illustrate the steps of the Ratio-to-Trend method described above.

Example 9.13: The manager of a company wants to analyze the quarterly demand of steering wheels to set quarterly production schedules. The following data showing quarterly sales records for 2004, 2005 and 2006 is available to the manager.

Quarter	2004	2005	2006
I	118	126	143
II	109	124	140
III	93	108	127
IV	120	139	155

- Estimate the trend equation.
- Compute the quarterly seasonal indices using the ratio-to-trend method.
- Explain how the manager can use these indices to set quarterly production schedules.

Solution:

Step 1: The unit in this case is a quarter.

Step 2: To compute the trend equation, we arrange the data as follows:

(1) Year	(2) Quarter (t)	(3) Sales (Y_t)	(4) Trend Values (Estimated) (T_t)	(5) % of Trend Values (Y_t/T_t)
2004	I	118	105.2432	112.1213
	II	109	108.8652	100.1238
	III	93	112.4872	82.67607
	IV	120	116.1092	103.351
2005	I	126	119.7312	105.2357
	II	124	123.3532	100.5243
	III	108	126.9752	85.0559
	IV	139	130.5972	106.4341
2006	I	143	134.2192	106.5421
	II	140	137.8412	101.5662
	III	127	141.4632	89.776
	IV	155	145.0852	106.8338

Using the method of least squares, the trend line is $\hat{Y}_t = 101.6212 + 3.622t$

Step 3: This equation is now used to estimate the trend values (T_t) for each quarter as shown in column (4) of table above.

Step 4: In column (5) of the table above, we next express each observed value as a percentage of the trend values i.e. we calculate $\left(\frac{Y_t}{T_t}\right) \times 100$

Step 5: We now arrange the ratios as follows to examine the presence or absence of seasonal variations:

Quarter	2004	2005	2006	Median (Seasonal Index)
I	112.1213	105.2357	106.5421	106.5421
II	100.1238	100.5243	101.5662	100.5243
III	82.67607	85.0559	89.776	85.0559
IV	103.351	106.4341	106.8338	106.4341

Step 6: Next the median for each quarter is calculated to eliminate the effect of irregular movements from the time series as given in the last column of the above table.

Step 7: The seasonal indices are now adjusted by multiplying each index by the adjustment factor calculated as follows:

$$\text{Adjustment Factor} = \frac{100}{99.6391} = 1.0036$$

The final seasonal indices (adjusted) are as follows:

Quarter	Unadjusted Seasonal Indices	Adjusted Seasonal Indices
I	106.5421	(106.5421) (1.0036) = 106.9256
II	100.5243	(100.5243) (1.0036) = 100.8862
III	85.0559	(85.0559) (1.0036) = 85.3621
IV	106.4341	(106.4341) (1.0036) = 106.8173

Managerial Implications: The first and the fourth quarter show a positive seasonal variation. The second quarter shows very marginal positive effect. However the third quarter definitely shows a negative seasonal effect indicating low sales. The manager can make use of these indices to schedule production to be higher in the first and the fourth quarter and have a lower production schedule in the third quarter.

9.5.3 Ratio-to-Moving Average Method

This method is similar to the ratio-to-trend method, the only difference being that in place of the trend values, moving averages are used and observed values are calculated as a percentage of the corresponding moving averages. A discussion of the steps of the Ratio-to-Moving Average method of calculating the seasonal indices follows:

Step 1: Calculate the appropriate moving average values. Center the values if necessary.

Step 2: Calculate each observed value (Y_t) as a percentage of the moving average values i.e.

compute the ratio $\left(\frac{Y_t}{\text{Moving Ave rage}}\right) \times 100$

Step 3: Identify if any seasonal variations are present by determining whether the period-to-period percentages are similar and also by observing whether the percentages of one period are distinctly different from atleast one more period.

Step 4: Next, as in the previous method, compute the medians or the modified means.

Step 5: Finally, the indices are adjusted to a average of 100 by multiplying each index by the following Adjustment Factor (AF):

$$\mathbf{AF} = \frac{\mathbf{100}}{\mathbf{\text{mean of the unadjusted indices}}}$$

These steps are now executed with the help of an example.

Example 9.14: Calculate the seasonal indices from the following data by the method of ratio-to-moving average method. This data is same as the one used to explain the concept of moving averages in section 9.4.3.

Year	I	II	III	IV
2004	42	58	80	60
2005	46	60	82	64
2006	44	56	85	70
2007	48	54	89	72

Solution:

Step 1: The moving average values already calculated are used directly.

Step 2: The observed values are calculated as a percentage of the moving average values as follows:

$\left(\frac{Y_t}{\text{Moving Ave rage}}\right) \times 100$ · These ratios are shown in column (5) of the table given below:

Year	Quarter	Centered Moving Average	Y_t	$\left(\frac{Y_t}{\text{Moving Average}}\right) \times 100$
2004	I			
	II			
	III	60.5	80	132.33
	IV	61.25	60	97.96
2005	I	61.75	46	74.49
	II	62.5	60	96
	III	62.75	82	130.68
	IV	62.00	64	103.68
2006	I	61.875	44	71.11
	II	63	56	88.89
	III	64.25	85	132.29
	IV	64.5	70	108.95
2007	I	64.75	48	74.13
	II	65.5	54	82.44
	III			
	IV			

Step 3: Identification

To identify whether any seasonal variation is present we now examine the period-to-period percentages

Quarter	2004	2005	2006	2007
I		74.49	71.11	74.13
II		96	88.89	82.74
III	132.23	130.68	132.29	
IV	97.96	103.68	108.95	

The percentages of quarter one show distinct similarity over all the three years 2004, 2005 and 2006. Also, the percentages corresponding to quarter III are similar for the years 2004, 2005 and 2006. Further, the percentages of quarter I are distinctly different from the percentages of quarter III, which in turn are quite distinctly different from those of quarter II. All these definitely indicate the presence of seasonal variations.

Step 4: Next, to smooth out the irregular variations, the medians (or alternatively the modified means) are computed, in this case.

Quarter	2004	2005	2005	2007	Median (Seasonal Index)
I		74.49	71.11	74.13	74.13
II		96	88.89	82.44	88.89
III	132.23	130.68	132.29		132.23
IV	97.96	103.23	108.95		103.23

Step 5: And finally the indices are adjusted to an average of 100, by multiplying each index by the following AF

$$AF = \frac{100}{\text{mean of the unadjusted indices}} = \frac{100}{99.62} = 1.0038$$

The final seasonal indices (adjusted) are shown in the following table.

Quarter	Unadjusted Seasonal Indices	Adjusted Seasonal Indices
I	74.13	(74.13) (1.0038) = 74.41
II	88.89	(88.89) (1.0038) = 89.23
III	132.23	(132.23) (1.0038) = 132.73
IV	103.23	(103.23) (1.0038) = 103.62

Interpretation

The seasonal index of quarter III definitely shows a very positive seasonal effect, while the seasonal index for the first quarter indicates a pronounced negative effect.

9.6 MEASUREMENT OF CYCLICAL COMPONENT

A number of methods are available for computation of cyclical variations, but most of them are beyond the scope of this book. However, the most commonly used method called the Residual Method is discussed here in brief.

9.6.1 Residual Method

Assuming that seasonal and irregular variations do not have any long term effect, except for irregular variations caused by major events like wars, famines, floods etc. this method reduces the multiplicative model of time series given by $Y_t = T_t \times S_t \times C_t \times I_t$ as follows:

$$Y_t = T_t \times C_t$$

Dividing this equation by T_t

$$Y_t = C_t$$

which gives an estimate of the cyclical component. Thus, the steps in the measurement of cyclical variation by this method are as follows:

Step 1: Using the annual data, the first step is to estimate the trend equation by using the method of least squares.

Step 2: The trend equation in step 1 is used to estimate all the trend values.

Step 3: The final cyclical indices are obtained by expressing the actual observed values as a percentage of the trend values i.e. $\left(\frac{Y_t}{T_t}\right) \times 100$

Interpretation

The interpretation is similar to that of the seasonal indices. A cyclical index greater than 100 indicates a positive cyclical variation and a cyclical index less than 100 indicates a negative cyclical variation.

We now discuss the computation of cyclical indices by the residual method with the help of an example.

Example 9.15: High rise construction limited is a construction company which builds high rise apartment buildings in major metro cities. The number of projects the company has undertaken for 20 years is shown in the following table.

High Rise Construction Projects 1985-2004

1985	15
1986	19
1987	22
1988	21
1989	19
1990	20
1991	21
1992	23
1993	26
1994	28
1995	27
1996	27
1997	25
1998	24
1999	26
2000	29
2001	34
2002	37
2003	39
2004	40

Calculate the cyclical indices by the Residual Method.

Solution:

Step 1: The trend equation is estimated by the method of least squares as follows:

$$T_t = \hat{Y}_t = 13.6315 + (1.13985)t$$

$t = 20$, for 2004

Step 2: Using the above equations we calculate the trend values for all the years as shown in column (3) of the table below

Step 3: And finally, the cyclical indices are computed by calculating the ratios $\left(\frac{Y_t}{T_t}\right) \times 100$ as shown in column 4 of the table below:

High Rise Construction Projects 1985-2004

(1) Year	(2) No. of Projects (Y_t)	(3) Estimated Trend (T_t)	(4) Cyclical Index $\left(\frac{\text{No. of Projects}}{\text{Estimated Trend}} \times 100\right)$
1985	15	14.77	101.55
1986	19	15.91	119.41
1987	22	17.05	129.02
1988	21	18.19	115.44
1989	19	19.33	98.29
1990	20	20.47	97.70
1991	21	21.61	97.17
1992	23	22.75	101.19
1993	26	23.89	108.83
1994	28	25.03	111.86
1995	27	26.17	103.17
1996	27	27.31	98.87
1997	25	28.45	87.87
1998	24	29.59	81.11
1999	26	30.73	84.61
2000	29	31.87	90.99
2001	34	33.01	103.00
2002	37	34.15	108.35
2003	39	35.29	110.52
2004	40	36.43	109.80

9.7 MEASUREMENT OF IRREGULAR COMPONENT

Since irregular variations are completely random in nature, they are difficult to model and analyze. The usual way to analyze it is to consider it as the leftover component once the seasonal, trend and cyclical component have been accounted for.

Also, often experts with knowledge about the behavior of the series may also be consulted to understand irregular variations.

9.8 BUSINESS FORECASTING: AN APPLICATION OF TIME SERIES ANALYSIS

To plan any activity, be it a new venture or a new product launch and so on, an industry or any business requires projections about future demand and supply to act as an indicator. Such decisions are often made in the face of uncertainty. Business forecasting involves a study of the internal and external forces that shape demand and supply. Business forecasting methods are broadly classified into two categories viz:

- (i) Qualitative Forecasting Methods
- (ii) Quantitative Forecasting Methods

(i) Qualitative Forecasting Methods

These are

- The Delphi Method
- The Sales Force Composite Method
- Consumer Panel Survey Method

The Delphi Method

The Delphi method is based on the opinions and experience of an expert panel in the particular area. The expert panel is also often referred to as a jury of expert opinion and the forecasts made are known as executive forecasts. Each member of the panel makes a knowledgeable forecast based on factual data and personal judgement. The entire data from the panel is then summarized to arrive at a consensus.

The Sales Force Composite Method

In this method, sales representatives are asked to make estimates of quantities of a product or a group of products expected to be sold in their respective areas. Such forecasts are called sales force estimates.

Consumer Panel Survey Method

A number of companies maintain a certain number of consumer panels consisting of a number of households. This method of forecasting uses a survey of such panels to arrive at a forecast.

- (ii) **Quantitative Forecasting Methods:** These methods are applicable when information can be quantified and historical data about the forecast variable is available. Time series analysis is a quantitative forecasting method, the other quantitative technique being regression models which has been discussed in Chapter 8.

The following chart shows the broad classification of business forecasting methods.

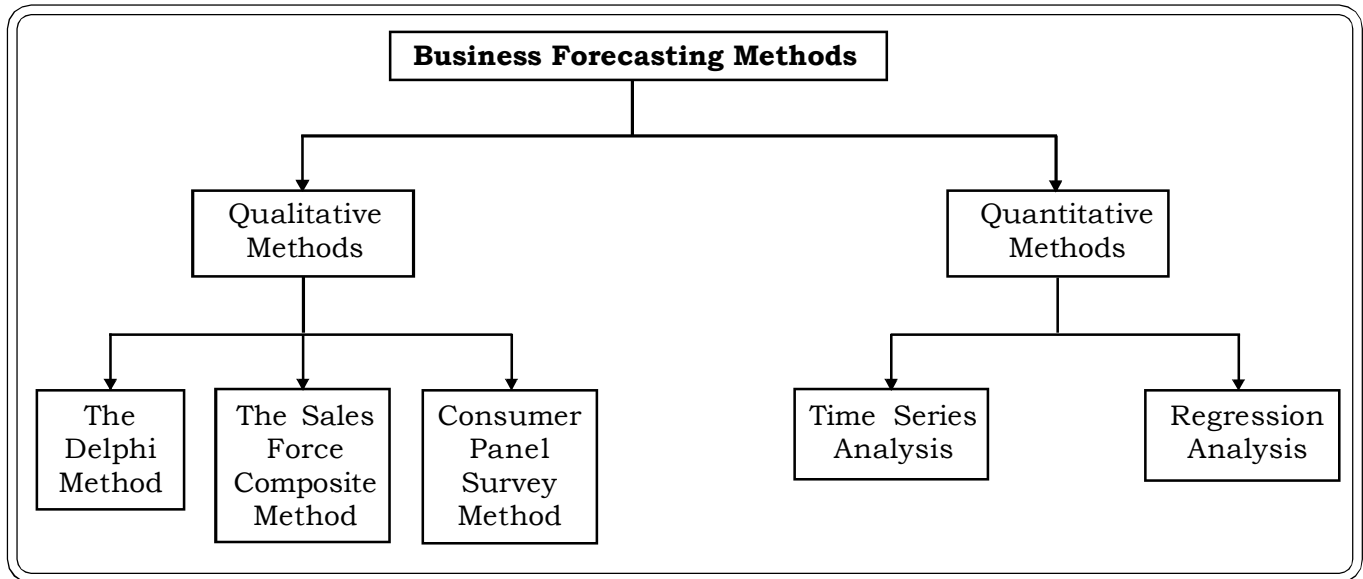


Figure 9.11

Business Forecasting Methods

The time series forecasting methods are:

- (i) Freehand method
- (ii) Smoothing methods
- (iii) Exponential Smoothing methods
- (iv) Trend adjusted for seasonal variation

The first two methods have been discussed in sections 9.4.1, 9.4.2 and 9.4.3.

This section presents a discussion of the exponential smoothing method and the trend adjusted for seasonal variation method.

9.8.1 The Exponential Smoothing Method

This is a popular method of forecasting and is a kind of weighted average that assigns more weights to the most recently observed value of the series and the most recent forecast. Thus, by this method.

$$\begin{aligned}\hat{Y}_{t+1} &= \alpha Y_t + (1-\alpha)\hat{Y}_t \\ &= \hat{Y}_t + \alpha(Y_t - \hat{Y}_t)\end{aligned}$$

where

\hat{Y}_{t+1} - is the forecast for the next or (t + 1) period.

\hat{Y}_t - is the forecast for the current period (t)

Y_t - observed value for period t

α - exponential smoothing constant

$$0 \leq \alpha \leq 1$$

A few points need consideration here:

(i) Choice of α

When the smoothing constant is close to 0, the series is not very responsive to current values. Therefore for an erratic series, a smaller value of α , close to zero is desirable.

Alternatively, when α is close to 1, the series is more responsive to current values. Thus, for a stable series, it is desirable to assign a value close to 1.

Alternatively, an appropriate value of the smoothing constant can be estimated as

$$\alpha = \frac{2}{n+1}$$

where n - the number of moving average period.

- (ii) A starting or first forecast of the series often called "seed". A usual practice is to take the first value in the time series as the first forecast value.
- (iii) If the series is subject to a pronounced trend or frequent oscillations, then this method is not very suitable. A two parameter method can be used in case of the presence of trend, as it takes into consideration the influence of trend. Further, if both trend and seasonal influences are present in the data, then a three parameter method is suitable.

Example 9.16: A hotel in a coastal town in India recorded the following room occupancies in the months of November and December.

Months	November	December
Room Occupancies	142	147

Using $\alpha = 0.20$ and $\alpha = 0.8$, what would be the forecast for the month of January using the exponential smoothing technique? The forecast for November was 145.

Solution:

The exponential smoothing model is

$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha(Y_t - \hat{Y}_t)$$

when $\alpha = 0.20$

Months	Room Occupancies (Actual)	Room Occupancies (Forecast)
November	142	145
December	147	$145 + 0.20(142 - 145) \cong 144$
January		$144 + 0.20(147 - 144) = 144.6 \cong 145$

Thus the forecast for the month of January is 145 room occupancies when $\alpha = 0.20$

The forecast for the month of December is as follows:

$$\hat{Y}_{t+1} = \text{Forecast for December}$$

$$\hat{Y}_t = \text{Forecast for November} = 145$$

$$Y_t = \text{Actual observed room occupancies in November} = 142$$

$$\text{December Forecast} = 145 + 0.20 (142 - 145) \cong 144$$

When $\alpha = 0.80$

Months	Room Occupancies (Actual)	Room Occupancies (Forecast)
November	142	145
December	147	$145 + 0.80 (142 - 145) = 142.6143 \cong 143$
January		$143 + 0.80 (147 - 143) = 146.2 \cong 146$

In this case, there is not much discrepancy in the two forecasts.

9.8.2 Trend adjusted for Seasonal Index Method

Forecast values using the trend equation can be improved considerably by combining the appropriate seasonal index. This can be done by the following method.

$$\text{New Forecast Value} = \frac{(\text{Estimated Trend Value}) \times (\text{appropriate seasonal index})}{100}$$

Such an adjustment can enhance the precision of the forecast to a large extent.

Example 9.17: Consider the data in example 9.12 related to quarterly demand of steering wheels to set quarterly production schedules.

Obtain new trend values adjusted for seasonal variations.

Solution:

The seasonal indices for the four quarters are:

Quarters	I	II	III	IV
Seasonal Index	106.5421	100.5243	85.0559	106.4341

The adjusted trend values are obtained as follows:

Year	Quarter	Trend Values (unadjusted)	Trend values (adjusted for seasonal variations)
2004	I	105.2432	$\frac{(105.2432 \times 106.5421)}{100} = 112.1238$
	II	108.8652	$\frac{(108.8652 \times 100.5243)}{100} = 109.4359$
	III	112.4872	$\frac{(112.4872 \times 85.0559)}{100} = 123.5798$
	IV	116.1092	$\frac{(116.1092 \times 106.4341)}{100} = 123.5798$
2005	I	119.7312	$\frac{(119.7312 \times 106.5421)}{100} = 127.5641$
	II	123.3532	$\frac{(123.3532 \times 100.5243)}{100} = 123.9999$
	III	126.9752	$\frac{(126.9752 \times 85.0559)}{100} = 107.9998$
	IV	1300.5972	$\frac{(130.5972 \times 106.4341)}{100} = 138.9999$
2006	I	134.2192	$\frac{(134.2192 \times 106.5421)}{100} = 142.9999$
	II	137.8412	$\frac{(137.8412 \times 100.5243)}{100} = 138.5639$
	III	141.4632	$\frac{(141.4632 \times 85.0559)}{100} = 120.3228$
	IV	145.0852	$\frac{(145.0852 \times 106.4341)}{100} = 154.4201$

9.9 CASELETS

1. Ahmedabad Jai Auto Engineering is one of the foremost manufacturing companies. The project leader of the R & D department of this company believes that the firms annual profits depend highly on the amount spend on R & D and suggests an increase in the R & D expenditure. However, the newly appointed Chief Executive Officer is hesitant to believe and asks for evidence.

The data of annual profit and R & D expenditure for past few years from the records is as follows.

Year	R & D expenditure (Rs crores)	Annual profit (Rs crores)
2002	2	26
2003	3	27
2004	5	35
2005	4	32
2006	11	44
2007	5	33

Q. How would you convince the Chief Executive to increase expenditure on R & D?

2. The growing integration of economies and societies around the world – has been one of the most widely debated topics in international economics over the past few years. Rapid growth and poverty reduction in China, India, and other countries that were poor 20 years ago, has been a positive aspect of globalization. But globalization has also generated significant international opposition over concerns that it has increased inequality and environmental degradation.

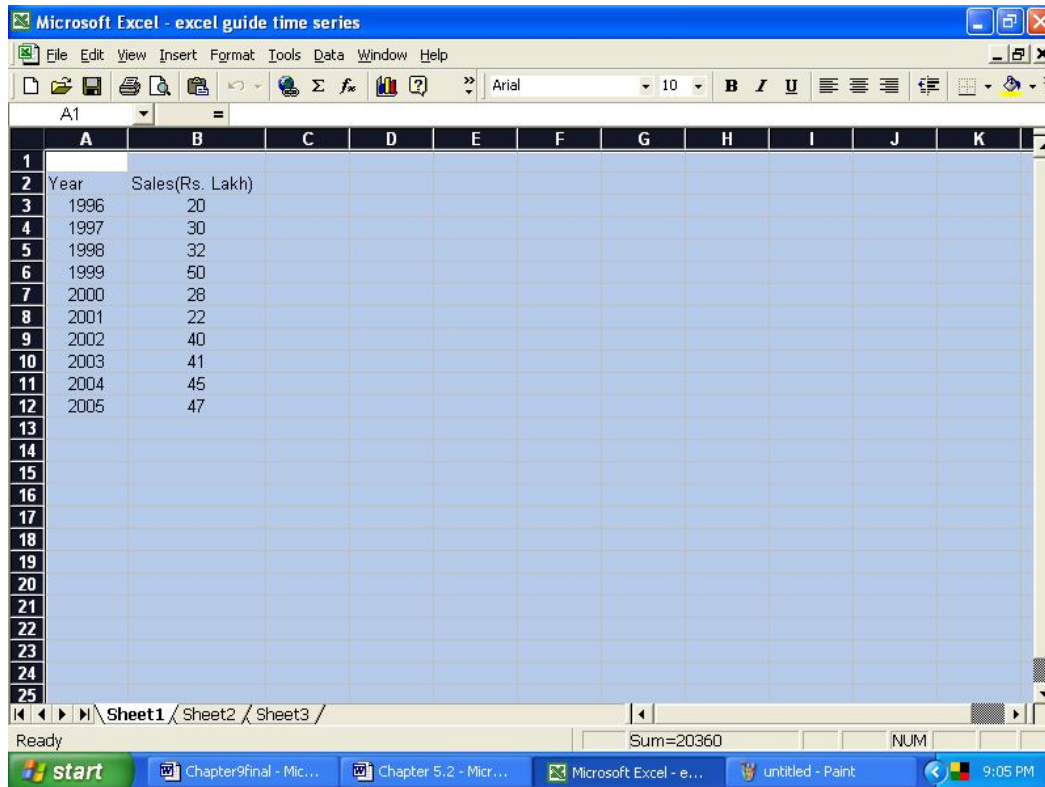
The following table shows the data of annual growth rate of Industrial sector of India for almost two decades:

Year	Annual growth rate of Industrial Sector (%)
1981	9.3
1982	3.2
1983	6.7
1984	8.7
1985	8.9
1986	9.1
1987	7.3
1988	8.7
1989	8.6
1990	8.2
1991	.6
1992	2.3
1993	6
1994	9.4
1995	12.4
1996	7.1

Q. Is there any favorable impact of Globalization in this sector?

9.10 EXCEL GUIDE**Calculation of Moving Average**

Step 1: Enter the data in an Excel spreadsheet as follows. (The data shown here is from Example 9.3)

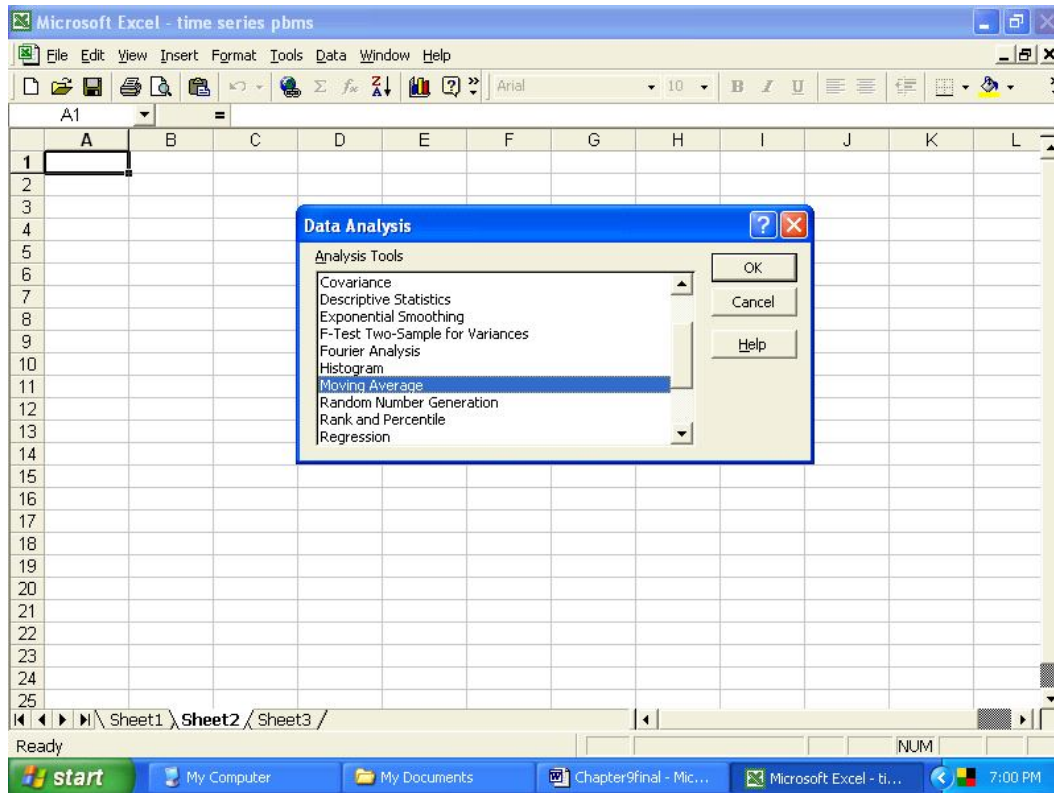


The screenshot shows a Microsoft Excel window titled "Microsoft Excel - excel guide time series". The spreadsheet contains the following data:

Year	Sales(Rs. Lakh)
1996	20
1997	30
1998	32
1999	50
2000	28
2001	22
2002	40
2003	41
2004	45
2005	47

The status bar at the bottom shows "Sum=20360" and "NUM". The taskbar at the bottom includes the Start button and several open applications: Chapter9final - Mic..., Chapter 5.2 - Mic..., Microsoft Excel - e..., and untitled - Paint. The system clock shows 9:05 PM.

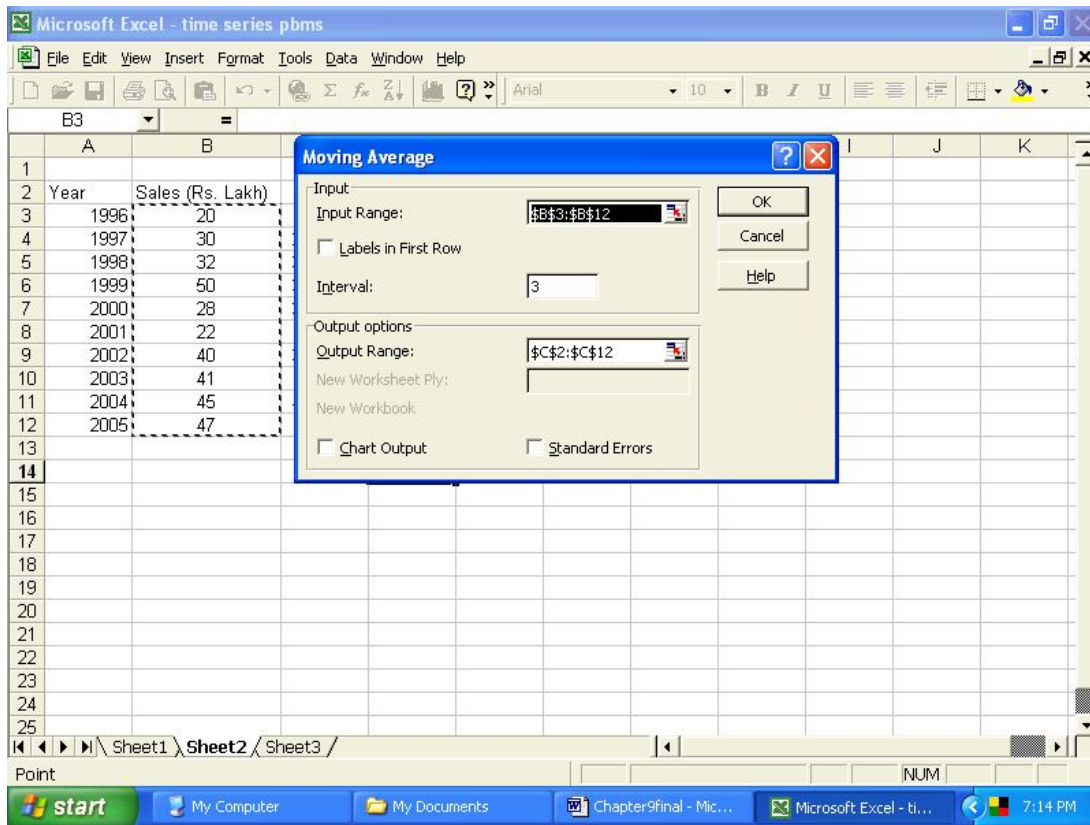
Step 2: Go to Tools in the menu bar and Select Data Analysis. From the Analysis tools, select Moving Average.



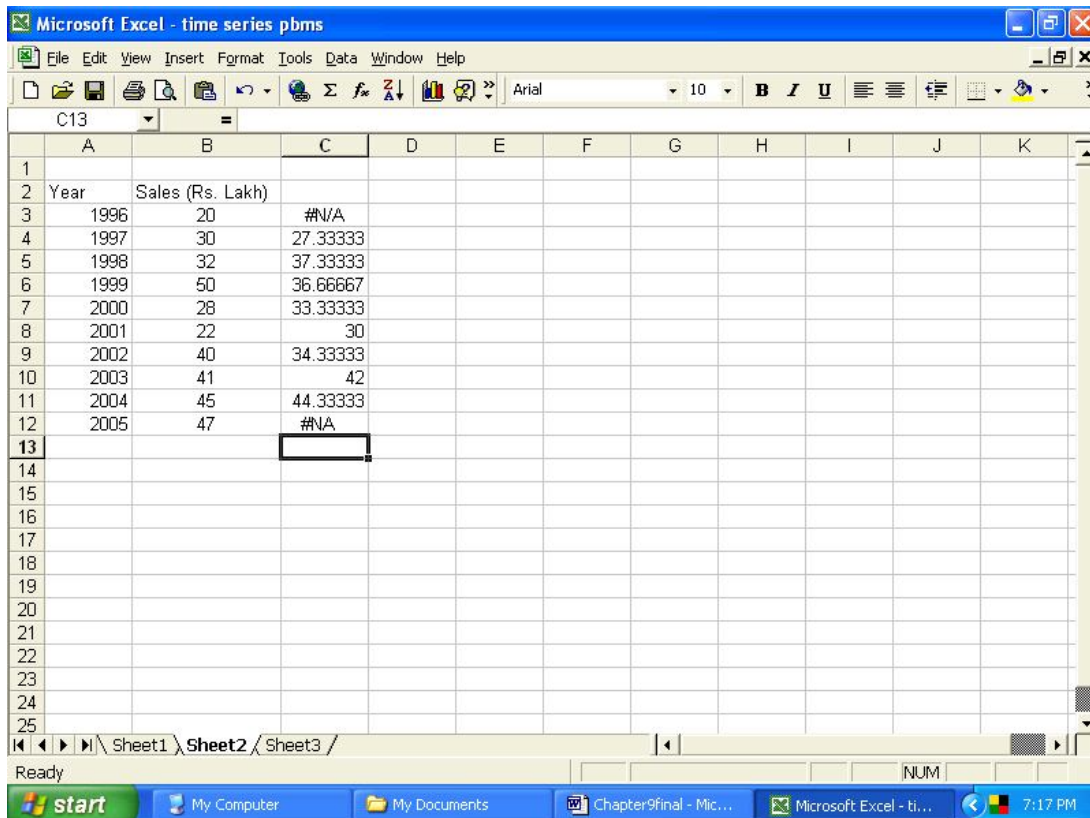
Step 3: Give the input range in this case it is B3:B12. In interval give the moving average time period. Here we need to calculate the three yearly moving average. So, the interval no is 3.

And finally give the output range. Here we have given it as C2: C12.

Finally click OK.

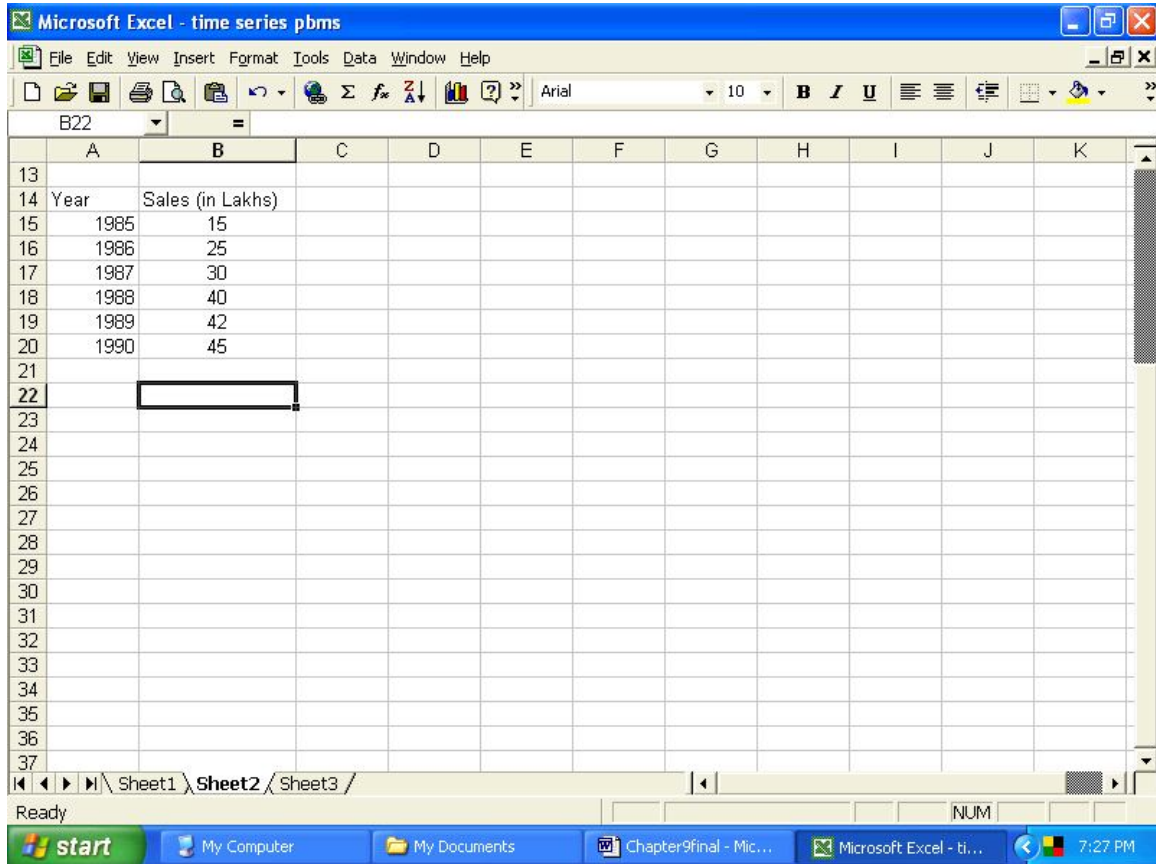


Step 4: The moving averages would be displayed as follows. Please note that the two moving averages that are not computable are given as Not Available (NA).



Calculation of Trend Values

Step 1: Enter the data. The data shown here is from example 9.8



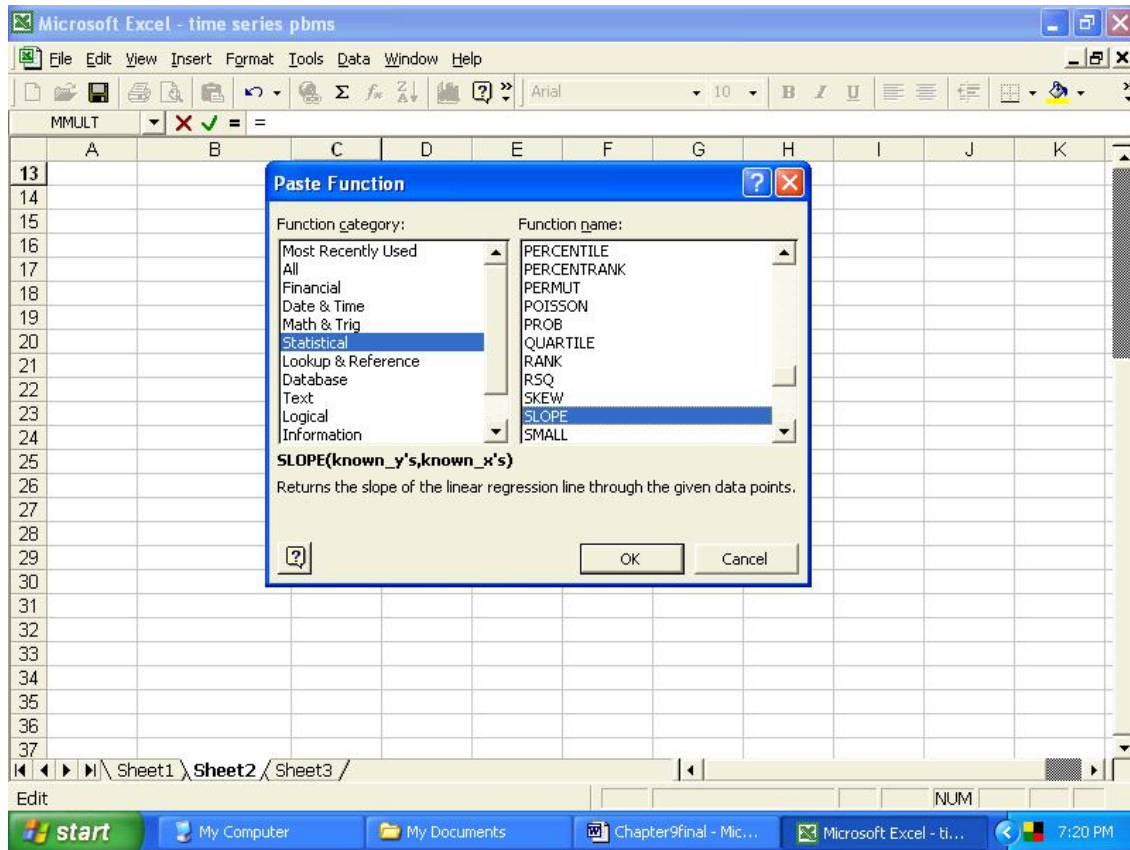
The screenshot shows a Microsoft Excel window titled "Microsoft Excel - time series pbms". The spreadsheet has the following data:

Year	Sales (in Lakhs)
1985	15
1986	25
1987	30
1988	40
1989	42
1990	45

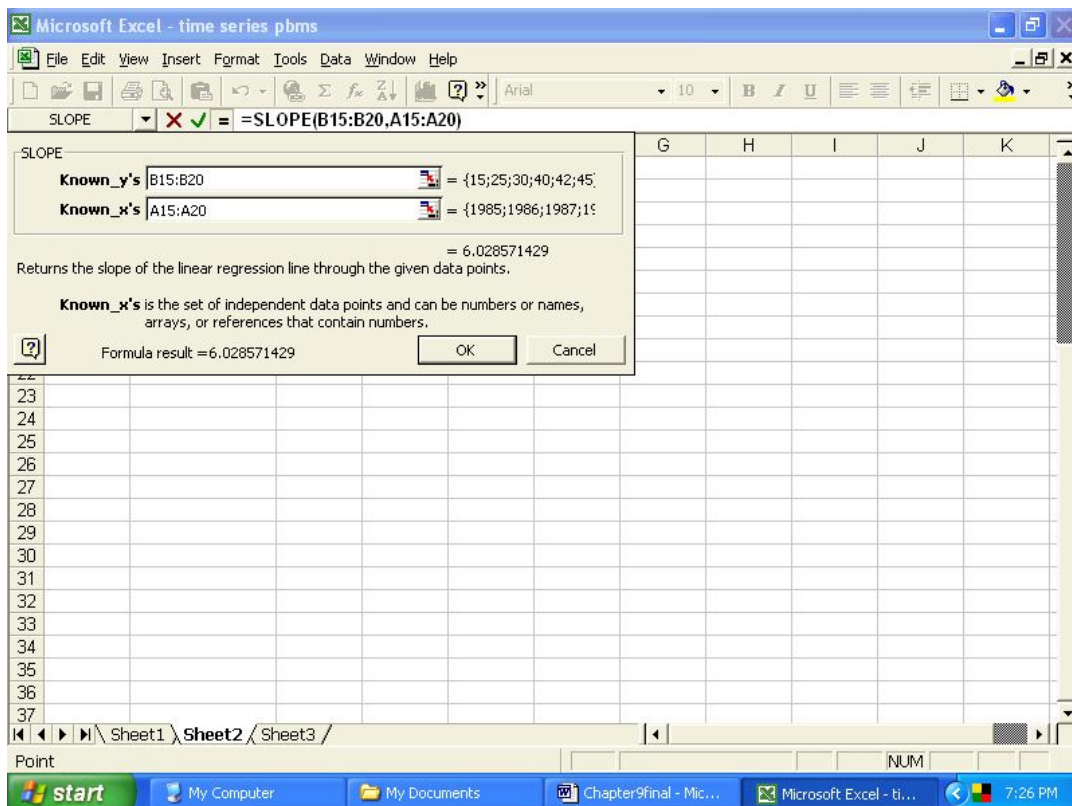
The formula bar is empty, and cell B22 is selected. The status bar at the bottom shows "Ready" and "NUM".

Step 2: From the formula bar (=), go to MORE FUNCTIONS.

Select STATISTICAL FUNCTIONS and from there select the function SLOPE.



Step 3: Enter the range for the X values and the Y values as shown below and click on OK.



Step 4: The final output would be displayed with the value of the slope .

Year	Sales (in Lakhs)
1985	15
1986	25
1987	30
1988	40
1989	42
1990	45

Slope =	6.028571429
---------	-------------

Step 5: To find the intercept, the steps are same. Only in Step 2, from STATISTICAL FUNCTIONS select the function INTERCEPT and the rest of the procedure remains the same.

Once the slope and the intercept are calculated, the trend line may be fitted.

9.11 EXERCISES

- 9.1 Define a time series with suitable examples.
- 9.2 How would you graphically display a time series? Explain with an example.
- 9.3 What are the four components of a time series? Explain briefly each component.
- 9.4 What are the models available for decomposition of a time series? State the situations in which each model is appropriate.
- 9.5 Describe the free hand curve fitting method. What are its limitations?
- 9.6 Explain how you would use the semi-average method to estimate trend.
- 9.7 Explain the term 'smoothing of a time series'. Name some popular methods of smoothing.
- 9.8 In moving average, what is 'centering'? Explain with an example.
- 9.9 When would you use the method of weighted moving average?
- 9.10 Define the exponential curve and state when it is used to estimate trend.

- 9.11 Name the methods available to estimate the seasonal component of a time series. Briefly describe one method.
- 9.12 How would you interpret seasonal indices?
- 9.13 Explain the Residual Method of estimating the cyclical component of a time series?
- 9.14 Describe the qualitative and quantitative techniques for business forecasting.
- 9.15 Explain the Exponential Smoothing Technique of Time Series Forecasting. Also elaborate on the significance of the Exponential Smoothing constant. State a situation when this technique is not suitable.
- 9.16 How would you adjust trend values for the effect of seasonal variations to improve forecasts?
- 9.17 A fast food chain has the following data related to the sale of burgers for the past twelve months since it opened. Fit a trend line by the Semi-Average Method.

Months	1	2	3	4	5	6	7	8	9	10	11	12
No. of Burgers sold (in 000)	20	22	24	25	28	29	31	30	32	37	39	41

Using this line, estimate the number of burgers expected to sell in the next month.

- 9.18 For the following data, find
- (i) the four yearly moving average values
- (ii) The three yearly moving average values

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Demand	28	34	42	94	40	25	42	68	39	29	48	82

- 9.19 The manager of a movie theatre wants to scheduler staffing and organize promotional schemes. She decides to analyze data for the past four weeks. Use a seven day moving average value to compute seasonal indices by using the ratio to moving average method.

Day	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Week 1
Attendance	175	196	199	205	410	562	370	
Day	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Week 2
Attendance	184	200	222	213	426	565	382	
Day	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Week 3
Attendance	187	215	210	235	440	595	385	
Day	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Week 4
Attendance	197	195	220	235	455	620	415	

9.20 Calculate the seasonal indices by the ratio-to-trend method for the following data.

Year	Quarter I	Quarter II	Quarter III	Quarter IV
2004	40	73	48	60
2005	38	70	50	58
2006	35	74	52	63



10

Chi - Square and Analysis of Variance



Structure

- 10.1 Introduction
- 10.2 The Chi - Square Statistic
 - 10.2.1 Chi-Square Test for Equality of Population Proportions or Chi - Square test for Homogeneity
 - 10.2.2 Chi- Square Test for Independence of Two Attributes
 - 10.2.3 Yates Correction for Continuity
 - 10.2.4 Chi- Square Test of Goodness of Fit
- 10.3 One way ANOVA
- 10.4 Assumptions of ANOVA
- 10.5 Simple Steps for ANOVA Calculations
- 10.6 Caselets
- 10.7 Excel Guide
- 10.8 Exercises

10.1 INTRODUCTION

In the chapter on estimation and testing we examined tests of hypothesis concerning a single population and two populations. For quantitative variables we defined tests for a single mean and difference of two means. And for variables whose nature was qualitative we defined tests for examining single proportion and difference of two proportions. However, in practice, often the case arises wherein we may have to test hypothesis regarding more than two populations. The hypothesis may concern qualitative as well as quantitative variables. In this chapter, we will discuss two important concepts related to testing of qualitative and quantitative variables when more than two populations are being studied. More specifically, for testing a hypothesis regarding equality of more than two population proportions we use a chi – square test and for testing a hypothesis regarding equality of more than two population means an Analysis of Variance (ANOVA) is suitable.

10.2 THE CHI - SQUARE STATISTIC

The chi – square statistic along with its distribution was defined in chapter 6.

Mathematically, the chi – square statistic is defined as:

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

where $O_i = i^{\text{th}}$ observed frequency

and $E_i = i^{\text{th}}$ expected frequency, $i = 1, 2, \dots, n$

The calculated χ^2 value is then compared with the tabulated value of χ^2 at a appropriate level of significance and degrees of freedom. The degrees of freedom may be calculated as follows:

(i) For a single sample Test

$$\text{d.f.} = k - 1$$

where k is the number of categories

(ii) For data in the form of a cross-classification table (known as contingency table) for various levels of two or more independent sample,

$$\text{d.f.} = (r - 1) (c - 1)$$

where r = number of rows

c = number of columns

This statistic has many applications depending on the hypothesis being tested, some of which are

- (i) Test for equality of population proportions.
- (ii) Test for independence of two attributes.
- (iii) Test of Goodness of fit.

10.2.1 Chi-Square Test for Equality of Population Proportions or Chi-Square Test for Homogeneity

The chi – square test can be applied if we want to test the hypothesis that several populations are homogenous will respect to some characteristic. For example, people in different age groups may

be asked to specify the kind of TV programmes they prefer to watch. We may then proceed to test whether the population is homogeneous with respect to the type of television programmes they like to watch.

As an example, consider testing whether the proportion of congress supporters in Assam are same as the proportion of congress supporters in U.P. We may take samples of say 300 & 400 voters from Assam and U.P. respectively. Suppose the samples yield the following results:

Table 10.1
Distribution of Voters

	Assam	U.P.	Total
Congress	200	280	480
BJP	100	120	220
Total	300	400	700

This table is also called a contingency table for equality of two proportions.

In general, a 2×2 contingency table is as follows:

Table 10.2
 2×2 Contingency Table

	Variable 1	Data Type I	Data Type II	Total
Variable 2	Category 1	a	b	a + b
	Category 2	c	d	c + d
Total		a + c	b + d	

STEP 1

We need to evaluate the null hypothesis

H_0 : The proportion of congress supporters in Assam is same as the proportion of congress supporters in U.P.

Against the alternative hypothesis

H_1 : The proportion of congress supporters in Assam is different from the proportion of congress supporters in U.P.

This hypothesis can be tested by using the Z – test for differences of two proportions.

Suppose now we need to generalize this situation to include two or more states, say Delhi & Maharashtra. Our null hypothesis would now become

$$H_0: P_1 = P_2 = P_3 = P_4$$

Where P_1 = Proportion of congress supporters in Assam.

P_2 = Proportion of congress supporters in U.P.

P_3 = Proportion of congress supporters in Delhi.

P_4 = Proportion of congress supporters in Maharashtra.

The alternative hypothesis would be

$$H_1 : P_1 \neq P_2 \neq P_3 \neq P_4$$

STEP 2

We choose a level of significance α , usually 5% i.e. $\alpha = 0.05$.

For testing purpose, we now consider simple random samples from each of these states. The sample sizes could be same or different. Suppose, the data obtained from the four states is as given in the table 10.3.

Table 10.3
Observed Frequencies

	Assam	U.P.	Delhi	Maharashtra	Total
Congress	200	280	150	300	930
BJP	100	120	100	50	370
Total	300	400	250	350	1300

This table gives the observed frequency of voters in the four states.

Let, for sample data,

p_1 = Observed proportion of congress supporters in Assam

p_2 = Observed proportion of congress supporters in U.P.

p_3 = Observed proportion of congress supporters in Delhi

p_4 = Observed proportion of congress supporters in Maharashtra

Using the information in table 10.3,

$$p_1 = 0.67$$

$$p_2 = 0.7$$

$$p_3 = 0.6$$

$$p_4 = 0.86$$

STEP 3

We now need to calculate the expected frequencies of voters in the four states if the null hypothesis were true.

The formula for calculating the observed frequencies is

$$\frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

For example, the expected frequency of congress voters in Assam will be given by:

$$\frac{300 \times 930}{1300} \cong 215$$

Likewise, the expected frequency of BJP voters in Assam is given by

$$\frac{300 \times 370}{1300} \cong 85$$

Similarly the rest of the expected frequencies are calculated in the following table no. 10.4

Table 10.4
Expected Frequencies

	Assam	U.P.	Delhi	Maharashtra	Total
Congress	215	286	179	250	930
BJP	85	114	71	100	370
Total	300	400	250	350	1300

We arrange the observed and expected frequencies for the different categories as follows:

Table 10.5
Calculation of χ^2 -Statistic

Categories	O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Assam / Congress	200	215	225	1.05
Assam / BJP	280	286	36	0.13
U.P. / Congress	150	179	841	4.70
U.P. / BJP	300	250	2500	10
Delhi / Congress	100	85	225	2.65
Delhi / BJP	120	114	36	0.32
Maharashtra / Congress	100	71	841	11.85
Maharashtra / BJP	50	100	2500	25
				$\chi^2 = 55.7$

STEP 4

The test statistic for testing H_0 is:

$$\begin{aligned} \chi^2 &= \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \\ &= \sum_{i=1}^8 \frac{(O_i - E_i)^2}{E_i} \\ &= 55.7 \end{aligned}$$

This statistic follows a chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom.

where r = no. of rows

c = no. of columns

This distribution has been described in chapter 6.

In this example, there are $(2 - 1)(4 - 1) = 3$ degrees of freedom. The rejection region lies on the right tail of the distribution.

STEP 5

The tabulated value of $\chi^2(3, 0.05) = 9.48$

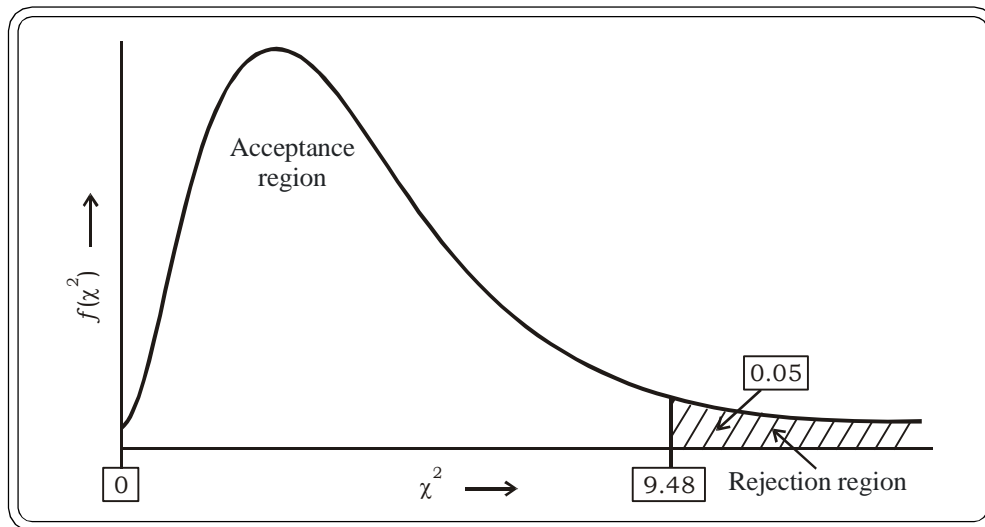


Figure 10.1

Graph of a χ^2 - distribution

STEP 6

Decision Rule:

The decision rule is to reject the null hypothesis if the calculated value of χ^2 comes to be greater than the tabulated value of χ^2 and to accept it otherwise. In this case, since calculated $\chi^2 >$ tabulated χ^2 we reject our null hypothesis.

STEP 7

Conclusion:

Voters in the four states are homogenous with respect to their political affiliations.

Steps in brief

Step 1: The first step is to define the null and the alternative hypothesis.

Step 2: Select a suitable level of significance.

Step 3: Then, the expected frequencies are calculated using the row-column rule for each cell.

Step 4: Calculate the chi-square test statistic viz.

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

where, O_i - i^{th} observed frequency (given)

E_i - i^{th} expected frequency (calculated in step 3)

Step 5: Calculate the tabulated value of the chi-square statistic. Compare it with the calculated value to arrive at a decision about accepting or rejecting the null hypothesis.

Usually, if $\text{cal } \chi^2 < \text{tab } \chi^2$, we may accept H_0 and reject otherwise.

Based on the decision, we final conclusion is given.

Remarks

Two points to be noted in chi – square tests are:

1. There are no assumptions made about the parent populations from which the samples have been drawn. In this respect the chi – square test is often regarded as a non-parametric test- i.e. a test that makes no assumptions about the population parameters.
2. The second point to be noted is that for the chi – square test to hold valid, none of the cell frequencies must be less than 5. In case any of the frequencies are less than 5, we have to make adjustments accordingly. (Discussed in section 10.2.3)

Example 10.1: For the following contingency table calculate the expected frequencies and the χ^2 statistic.

Category	A	B	C	Total
I	60	120	20	200
II	40	30	60	130
Total	100	150	80	330

Solution:

Expected Frequencies are:

Category	A	B	C	Total
I	61	91	48	200
II	39	59	32	130
Total	100	150	80	330

These have been calculated by using the formula:

$$\frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}, \text{ for each cell}$$

For example, let E_{ij} = expected frequency corresponding to the i^{th} row and the j^{th} column

R_i – i^{th} row total and C_j – j^{th} column total.

Then

$$E_{11} = \frac{R_1 \times C_1}{GT} = \frac{100 \times 200}{330} \cong 61$$

$$E_{12} = \frac{R_1 \times C_2}{GT} = \frac{150 \times 200}{330} \cong 91$$

$$E_{13} = \frac{R_1 \times C_3}{GT} = \frac{200 \times 80}{330} \cong 48$$

$$E_{21} = \frac{R_2 \times C_1}{GT} = \frac{130 \times 100}{330} \cong 39$$

$$E_{22} = \frac{R_2 \times C_2}{GT} = \frac{130 \times 150}{330} \cong 59$$

$$E_{23} = \frac{R_2 \times C_3}{GT} = \frac{130 \times 80}{330} \cong 32$$

The chi - square statistic can be conveniently calculated by arranging the data according to the different categories, in the form of a table, as follows:

Table 10.8

Categories	O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
A / I	60	61	1	0.164
A / II	40	39	1	0.026
B / I	120	91	841	9.242
B / II	30	59	841	14.254
C / I	20	48	784	16.33
C / II	60	32	784	24.5
				$\chi^2 = 64.52$

Thus, χ^2 - statistic = 64.52.

Example 10.2: A marketing research department of a television manufacturing company has chosen seven cities. It is believed that each city has the same sales potential. The actual number of television sets sold by the company in each city, during a one - month period, is given in the following table. Test the hypothesis that the seven cities have equal sales potential, using a level of significance of 0.05.

City	Number of sets sold
A	120
B	185
C	260
D	190
E	210
F	175
G	260
Total	1400

Solution:

The null hypothesis

H_0 : The seven cities have equal sales potential.

The alternative hypothesis

H_1 : The seven cities have different sales potential.

The test statistic

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \sim \chi_{(n-1)}^2$$

O_i = Observed sales in the i^{th} city.

E_i = Expected sales in the i^{th} city.

If the null hypothesis were true, the number of sets sold in each state is expected to be 200. Thus we get the following table:

City	(O_i)	(E_i)	($O_i - E_i$) ²	($O_i - E_i$) ² / E_i
A	120	200	6400	35
B	185	200	225	1.125
C	260	200	3600	18
D	190	200	100	0.5
E	210	200	100	0.5
F	175	200	625	3.125
G	260	200	3600	18
	1400	1400		$\chi^2 = 76.25$

Thus, calculated $\chi^2 = 76.25$

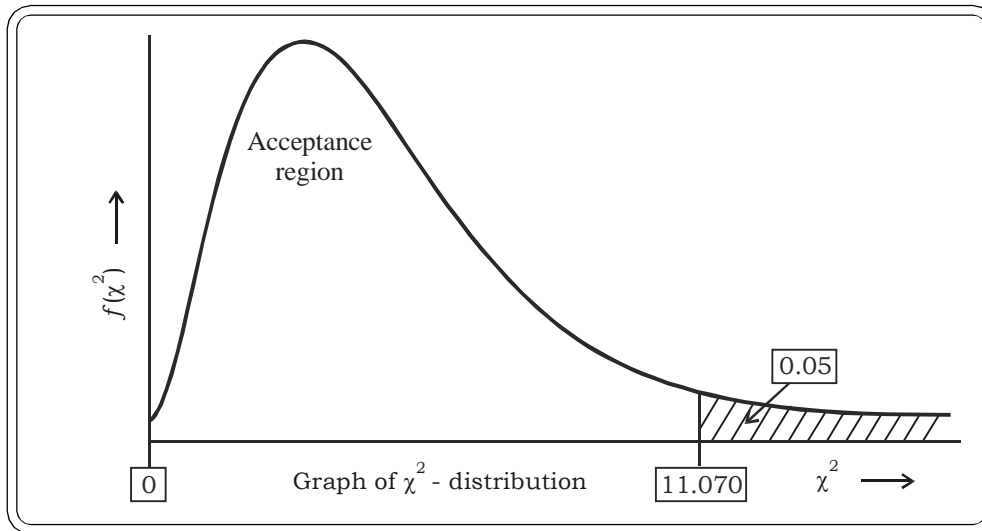
Tabulated $\chi^2 (.05, 6) = 12.592$, where

Degrees of freedom

$$= k - 1$$

$$= 7 - 1 = 6$$

since this is a one sample test.



Decision:

The null hypothesis is rejected at 5% level of significance since cal $\chi^2 >$ tabulated χ^2 .

Conclusion:

The sales potential of television sets sold is different across the seven cities.

Example 10.3: A newspaper conducted a telephonic poll on the issue of reservations for minorities in professional educational institutes. They wanted to find out if students in Delhi, Mumbai, Chennai and Kolkata have similar sentiments regarding minority reservation issue.

The following sample data containing observed frequencies was obtained:

	Delhi	Mumbai	Chennai	Kolkata	Total
In Favor	30	20	40	20	110
Against	70	80	60	80	290
Total	100	100	100	100	400

Use a 0.05 level of significance to test the hypothesis that the sentiment of students against reservation is the same in all the four cities.

Solution:**The null hypothesis**

H_0 : Proportion of students against reservation in professional educational institutes is same across the four cities surveyed.

The alternative hypothesis

H_1 : Proportion of students against reservation is different across the four cities i.e. sentiments against reservation differs in the four cities.

We now need to calculate the expected frequencies under the null hypothesis. This is done by using the formula

$$\frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}, \text{ for each cell.}$$

From the table of observed frequencies, the expected frequencies are calculated as follows:

	Delhi	Mumbai	Chennai	Kolkata	Total
In Favor	$\frac{110 \times 100}{400}$	$\frac{110 \times 100}{400}$	$\frac{110 \times 100}{400}$	$\frac{110 \times 100}{400}$	110
Against	$\frac{100 \times 290}{400}$	$\frac{100 \times 290}{400}$	$\frac{100 \times 290}{400}$	$\frac{100 \times 290}{400}$	290
Total	100	100	100	100	400

The final expected frequencies are:

	Delhi	Mumbai	Chennai	Kolkata	Total
In Favor	27.5	27.5	27.5	27.5	110
Against	72.5	72.5	72.5	72.5	290
Total	100	100	100	100	400

The test statistic

$$\chi^2 = \sum_{i=1}^8 \frac{(O_i - E_i)^2}{E_i}$$

For simplification of calculations we can arrange the data in the following format:

Category	Observed Frequencies (O_i)	Expected Frequencies (E_i)	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Delhi/ Favor	30	27.5	6.25	0.23
Mumbai/ Favor	20	27.5	56.25	2.05
Chennai / Favor	40	27.5	156.25	5.68
Kolkata / Favor	20	27.5	56.25	2.01
Delhi/ Against	70	72.5	6.25	0.09
Mumbai/ Against	80	72.5	56.25	0.78
Chennai / Against	60	72.5	156.25	2.16
Kolkata / Against	80	72.5	56.25	0.78
				$\chi^2 = 13.78$

Thus, $\chi^2 = 13.78$.

Degrees of freedom = $(r - 1)(c - 1) = (2 - 1)(4 - 1) = 3$

The tabulated value of $\chi^2(0.05, 3) = 7.815$ (from chi - square tables).

Decision:

Since the calculated χ^2 is greater than the tabulated value of χ^2 , we may reject the null hypothesis and accept the alternative hypothesis at 5% level of significance.

Conclusion:

Student sentiments for and against the issue of reservation seems to differ across the four cities Delhi, Mumbai, Chennai and Kolkata.

Example 10.4: A survey was conducted among student mobile phone users of four countries to determine if students in these countries downloaded the latest ring tones regularly on their mobile phones in equal proportion. Among 1000 students surveyed 60% Indians, 60% Chinese, 40% Americans and 40% British students answered in affirmative when asked if they downloaded ring tones regularly. The data is given in the following contingency table:

	Indian	Chinese	American	British	Total
Download regularly	400	360	240	160	1160
Do not download regularly	100	240	360	240	940
Total	500	600	600	400	2100

At 0.05 level of significance, determine whether there is significant difference in the proportion of Indian, Chinese, American and British students who downloaded the latest ring tones on their mobile phones. (Given data is hypothetical)

Solution:

Let P_1 = Proportion of Indian students who downloaded ring tones regularly.

P_2 = Proportion of Chinese students who downloaded ring tones regularly.

P_3 = Proportion of American students who downloaded ring tones regularly

P_4 = Proportion of British students who downloaded ring tones regularly.

The null hypothesis

$H_0: P_1 = P_2 = P_3 = P_4$ i.e. there is no significant difference in the proportion of Indian, Chinese, American & British Students who download ring tones regularly on their mobile phones.

The alternative hypothesis

$H_1: P_1 \neq P_2 \neq P_3 \neq P_4$

i.e. there is significant difference among the proportion of Indian, Chinese, American and British students who download ring tones regularly on their mobile phones.

The Expected Frequencies

	Indian	Chinese	American	British	Total
Down load regulary	$\frac{500 \times 1159}{2100}$ = 276	$\frac{600 \times 1159}{2100}$ = 331	$\frac{600 \times 1159}{2100}$ = 331	$\frac{400 \times 1159}{2100}$ = 221	1159
Do not download regulary	$\frac{500 \times 940}{2100}$ = 224	$\frac{600 \times 940}{2100}$ = 269	$\frac{600 \times 940}{2100}$ = 269	$\frac{400 \times 940}{2100}$ = 179	941
Total	500	600	600	400	2100

The test statistic

For calculating the χ^2 test statistic we make the following table:

Categories	(O _i)	(E _i)	(O _i - E _i) ²	$\frac{(O_i - E_i)^2}{E_i}$
Download/ Indian	400	276	15376	55.71
Not Download/ Indian	100	224	15376	68.64
Download/ Chinese	360	331	841	2.54
Not Download/ Chinese	240	269	841	3.13
Download/ American	240	331	8281	25.01
Not Download/ American	360	269	8281	30.78
Download/ British	160	221	3721	16.84
Not Download/ British	240	179	3721	20.79
Total	2100	2100		$\chi^2 = 223.44$

Calculated $\chi^2 = 223.44$

Tabulated $\chi^2 (0.05,3) = 7.815$

Degrees of freedom = $(r - 1) (c - 1) = (2 - 1) (4 - 1) = 3$

Decision:

The decision is to reject the null hypothesis, since calculated χ^2 is significantly greater than tabulated χ^2 , at 5% level of significance.

Conclusion:

There is significant difference among the proportion of Indian, Chinese, American & British students who download the latest ring tones on their mobiles.

Example 10.5: 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:

	No. of students in each level				
	Below Average (BA)	Average (A)	Alone Average (AA)	Genius (G)	
X	86	60	44	10	200
Y	40	33	25	2	100
126	93	69	12	300	

Would you say that the sampling techniques adopted by the two researchers are significantly different? (MBA, DU, 2001)

Solution:

The null & the alternative hypothesis.

H_0 : The sampling techniques adopted by the two researchers are not significantly different.

H_1 : The sampling techniques adopted by the two researchers are significantly different.

The expected frequencies:

The expected frequencies (E_i) are calculated by using the Row-Column Rule.

	No. of students in each level				
	Below Average (BA)	Average (A)	Alone Average (AA)	Genius (G)	
X	84	62	46	8	200
Y	42	31	23	4	100
126	93	69	12	300	

The Test Statistic

We make the following table for calculating the chi - square statistic.

Categories	(O_i)	(E_i)	($O_i - E_i$) ²	$\frac{(O_i - E_i)^2}{E_i}$
X/ BA	86	84	4	0.05
X/ A	60	62	4	0.05
X/ AA	44	46	4	0.09
X/ G	10	8	4	0.5
Y/ BA	40	42	4	0.09
Y/ A	33	31	4	0.13
Y/ AA	25	23	4	0.17
Y/ G	2	4	4	1
Total				$\chi^2 = 2.09$

Thus, calculated $\chi^2 = 2.09$.

Degrees of freedom = $(r - 1)(c - 1) = (2 - 1)(4 - 1) = 3$.

Level of significance = 0.05.

Tabulated $\chi^2(0.05, 3) = 7.815$.

Decision:

Since calculated χ^2 is greater than tabulated χ^2 , the null hypothesis is rejected.

Conclusion:

The sampling techniques adopted by the two researchers are significantly different.

Remark:

It may be noted that cell frequency corresponding to the last category is less than 5. In such cases we need to pool it with the previous cell frequency and then calculate the χ^2 - Test statistic. The procedure is described later in this chapter.

Example 10.6: Three pain relievers P_1 , P_2 & P_3 were tried on 100 patients. The extents of relief recorded by these patients are categorized as Excellent, Moderate and Poor. The results of the trial are given in the following 3×3 contingency table.

	P_1	P_2	P_3
Excellent	43	51	52
Moderate	37	28	20
Poor	20	21	28

At 5 percent level of significance, test whether the three pain relievers has the same degree of effectiveness.

Solution:

The null hypothesis

H_0 : The three types of pain relievers have the same degree of effectiveness.

The alternative hypothesis

H_1 : The three types of pain relievers have different degrees of effectiveness.

The observed frequencies (O_i):

	P_1	P_2	P_3	Row Totals
Excellent	43	51	52	146
Moderate	37	28	20	85
Poor	20	21	28	69
Column Totals	100	100	100	300

The expected frequencies (E_i):

	P_1	P_2	P_3	Row Totals
Excellent	49	49	49	147
Moderate	28	28	28	84
Poor	23	23	23	69
Column Totals	100	100	100	300

The χ^2 - test statistic

$$\begin{aligned}\chi^2 &= \frac{(49 - 43)^2}{49} + \frac{(51 - 49)^2}{49} + \frac{(52 - 49)^2}{49} + \frac{(37 - 28)^2}{28} + \frac{(28 - 28)^2}{28} + \frac{(20 - 28)^2}{28} \\ &\quad + \frac{(20 - 23)^2}{23} + \frac{(21 - 23)^2}{23} + \frac{(28 - 23)^2}{23} \\ &= 0.73 + 0.08 + 0.18 + 2.89 + 0 + 2.29 + 0.39 + 0.17 + 1.09 \\ &= 7.82\end{aligned}$$

Degrees of freedom = $(r - 1)(c - 1) = (3 - 1)(3 - 1) = 4$

Tabulated χ^2 (.05, 4) = 9.488

Decision:

Since calculated χ^2 is less than the tabulated χ^2 we may accept the null hypothesis, at 5% level of significance.

Conclusion:

The three types of pain relievers have the same level of effectiveness.

Example 10.7: Four different drugs have been developed for a certain disease. These drugs are used under three different environment (it is assumed that the environment might affect efficacy of drugs). The number of cases of recovery from the disease of people who have taken the drugs is tabulated as follows:

	Drugs				Row Totals
	A_1	A_2	A_3	A_4	
I	17	10	17	14	58
II	9	8	11	9	37
III	14	9	20	7	50
	40	27	48	30	145

Test whether the drugs differ in their efficiency to treat the disease.

Solution:

The null and alternative hypotheses are

H_0 : The drugs do not differ in their efficacy to treat the disease.

H_1 : The drugs differ in their efficacy to treat the disease.

Expected Frequencies

The expected frequencies are calculated as follows:

	A_1	A_2	A_3	A_4	Row Total
I	$\frac{40 \times 58}{145}$ = 16	$\frac{27 \times 58}{145}$ = 11	$\frac{48 \times 58}{145}$ = 19	$\frac{30 \times 58}{145}$ = 12	58
II	$\frac{40 \times 37}{145}$ = 10	$\frac{27 \times 37}{145}$ = 7	$\frac{48 \times 37}{145}$ = 12	$\frac{30 \times 37}{145}$ = 8	37
III	$\frac{40 \times 50}{145}$ = 14	$\frac{27 \times 50}{145}$ = 9	$\frac{48 \times 50}{145}$ = 17	$\frac{30 \times 50}{145}$ = 10	50
Column Total	40	27	48	30	145

The Test Statistic

For calculating the test statistic, we make a table as follows:

Categories	O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
I/A ₁	17	16	1	0.0625
I/A ₂	10	11	1	0.0909
I/A ₃	17	19	4	0.210
I/A ₄	14	12	4	0.333
II/A ₁	9	10	1	0.1
II/A ₂	8	7	1	0.143
II/A ₃	11	12	1	0.083
II/A ₄	9	8	1	0.125
III/A ₁	14	14	0	0
III/A ₂	9	9	0	0
III/A ₃	20	17	4	0.235
III/A ₄	7	10	9	0.9
				$\chi^2 = 2.2824$

Calculated $\chi^2 = 2.2824$

Degrees of freedom = $(r - 1)(c - 1) = (3 - 1)(4 - 1) = 2 \times 3 = 6$

Tabulated $\chi^2(6, 0.05) = 12.592$

Decision:

Since calculated χ^2 is less than tabulated χ^2 , we may accept the null hypothesis, at 5% level of significance.

Conclusion:

The drugs do not differ in their efficiency to treat the disease.

Example 10.8: A random sample of 150 families each showed the following number of girl child.

Girl Child	No. of families
0	30
1	30
2	40
3	25
4	25

Test the hypothesis that female and male births are equally likely to occur.

Solution:

The null and alternative hypothesis

H_0 : Female and male births are equally likely to occur.

H_1 : Female and male births are not equally likely to occur.

The test statistic

Under the null hypothesis, the expected number of female births would be equal.

Girl Child	No. of Families (O_i)	Expected (E_i)	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	30	30	0	0
1	30	30	0	0
2	40	30	100	3.33
3	25	30	25	0.83
4	25	30	25	0.83
				$\chi^2 = 4.99$

Thus, calculated $\chi^2 = 4.99$

Tabulated $\chi^2 (0.05, 4) = 9.49$

Decision:

Since calculated χ^2 is less than tabulated χ^2 , we may accept the null hypothesis.

Conclusion:

Female and male births are equally likely to occur.

10.2.2 Chi-Square Test for Independence of Two Attributes

A chi – square test can also be used to test independence of two attributes or two categorical variables. In this case there are two factors to be studied, each at different levels. Rows represent different levels of one factor and column represents different levels of the second factor. In testing for difference of proportions we had one factor of interest with two or more than two levels. The contingency table in this case is as follows.

Suppose we have factor A to be studied at n different levels and factor B at m different levels. The $m \times n$ contingency table will be:

Table 10.6
A $m \times n$ Contingency Table

	Factor A	Level 1	Level 2	Level n	Row Totals
Factor B	Level 1	O_{11}	O_{12}		O_{1n}	R_1
	Level 2	O_{21}	O_{22}		O_{2n}	R_2
	⋮					
	Level m	O_{m1}	O_{m2}		O_{mn}	R_m
	Column Totals	C_1	C_2		C_n	N

where O_{ij} – observed frequency in the i^{th} row and j^{th} column.

The same row-column rule is used to find the expected frequencies i.e.

E_{ij} – Expected frequency in the i^{th} row and j^{th} column.

$$= \frac{R_i \times C_j}{N}$$

where R_i = Total of the i^{th} row

C_j = Total of the j^{th} column

The chi – square statistic is:

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

where r = number of rows, c = number of columns

The null hypothesis in this case is usually of the following form

H_0 : The two attributes are independent

The alternative hypothesis is of the form

H_1 : The two attributes are not independent.

Once the expected frequencies are calculated, we compute the value of χ^2 and compare it with the tabulated value of χ^2 at a suitable level of significance.

The decision rule remains the same.

Example 10.9: A study was conducted on a sample of 500 men and women. The purpose was to study if there is any relationship between gender and the type of music they prefer. The following contingency table give the survey results.

Music Preferred	Male	Female	Row Totals
Classical	90	100	190
Popular Music	210	100	310
Column Total	300	200	500

At 5% level of significance test if gender and musical tastes are independent attributes.

Solution:

Null hypothesis

H_0 : Gender and musical tastes are independent.

Alternative hypothesis:

H_1 : Gender and musical tastes are dependent.

Expected frequencies

We now calculate the expected frequencies (E_{ij}) as follows:

Let E_{ij} = Expected frequency in the i^{th} row and the j^{th} column.

$$= \frac{R_i \times C_j}{GT}$$

where R_i = Total of the i^{th} row

C_j = Total of the j^{th} column.

GT = Grand Total

$$\text{Here, } E_{11} = \frac{R_1 \times C_1}{GT} = \frac{190 \times 300}{500} = 114$$

$$E_{12} = \frac{R_1 \times C_2}{GT} = \frac{190 \times 200}{500} = 76$$

$$E_{21} = \frac{R_2 \times C_1}{GT} = \frac{310 \times 300}{500} = 186$$

$$E_{22} = \frac{R_2 \times C_2}{GT} = \frac{310 \times 200}{500} = 124$$

We finally get the following total of expected frequencies.

Gender	Male	Female	Row Totals
Music Preferred			
Classical	114	76	190
Popular Music	186	124	310
Column Totals	300	200	500

The test statistic

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

The following table is made to calculate the test statistic

Categories	Expected Frequencies (E_i)	Observed Frequencies (O_i)	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
Classical / Male	114	90	576	5.05
Classical / Female	76	100	576	7.58
Popular Music / Male	186	210	576	3.10
Popular Music / Female	124	100	576	4.65
Total	500	500		$\chi^2 = 20.38$

Thus $\chi^2 = 20.38$

Tabulated $\chi^2 (.05, 1) = 3.841$ (from χ^2 tables)

Decision:

The tabulated χ^2 is much smaller than the calculated χ^2 . So there is not enough evidence to accept the null hypothesis and hence we may accept the alternative hypothesis.

Conclusion:

There is enough statistical evidence to conclude that there is a relationship between gender of a person and his/her taste in music.

Example 10.10: Four major automobile manufacturers in the Indian market are Maruti, Hyundai, Tata Motors and Ford. These companies manufacture cars in three segments the small car segment, the medium car segment and the large car segment. Suppose in a hypothetical study conducted to analyze whether there was any relationship between the manufacturer and the type of car preferred by consumers, the following data was obtained.

Car Type ↓	Maruti	Hyundai	Tata Motors	Ford	Row Totals
Small	80	60	75	10	225
Medium	15	10	15	30	70
Large	5	30	10	60	105
Column Totals	100	100	100	100	400

Test if there is any significant relationship between the manufacturer and the car type preferred by consumers.

Solution:

The null hypothesis

H_0 : The car manufacturer and the car type are independent.

The alternative hypothesis

H_1 : The car manufacturer and the car type are related.

The test statistic

Categories	Observed Frequencies (O_i)	Expected Frequencies (E_i)	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
Small/ Maruti	80	56.25	564	10
Medium/ Maruti	15	17.5	6.25	0.36
Large/ Maruti	5	26.25	451.5	17.2
Small/ Hyundai	60	56.25	14.06	0.25
Medium/ Hyundai	10	17.5	56.25	3.21
Large/ Hyundai	30	26.25	14.0625	0.54
Small/ Tata	75	56.25	351.56	6.25
Medium/ Tata	15	17.5	6.25	0.36
Large/ Tata	10	26.25	264	10.06
Small/ Ford	10	56.25	2139	38.03
Medium/ Ford	30	17.5	156.25	8.43
Large/ Ford	60	26.25	1139	43.39
				$\chi^2 = 138.58$

Calculate $\chi^2 = 138.58$

Tabulated $\chi^2 (.05, 6) = 12.592$

Decision:

Calculated value of χ^2 is much higher than tabulated value. We may reject the null hypothesis and accept the alternative hypothesis, at 5% level of significance.

Conclusion:

Based on the hypothetical data, there is statistical evidence to indicate that there seems to be a relationship between the car manufacturer and the type of car of a particular manufacturer, which a consumer prefers. For example Maruti may be the preferred brand in the small segment but of course such a hypothesis has to be further tested.

Example 10.11: A factory operates in three shifts. The following table gives the number of good and defective parts produced by each of the three shifts in the factory.

Shift	Good parts	Defective parts	Total
Day	900	130	1030
Evening	700	170	870
Night	400	200	600
	2000	500	2500

Is there any association between the shift and the quality of parts produced? [χ^2 (.05,2) = 5.991]

Solution:**The null hypothesis**

H_0 : There is no association between the shift and the quality of parts produced

The alternative hypothesis

H_1 : There is association between the shift and the quality of parts produced.

The test statistic

We need to find out the expected frequencies. This is done by using the row/ column formula and the following table is obtained.

Shift	Good	Defective	Row Totals
Day	824	206	1030
Evening	696	174	870
Night	480	120	600
Column Totals	2000	500	2500

For calculation of the χ^2 test statistic we now arrange the data as follows:

Categories	Observed Frequencies (O_i)	Expected Frequencies (E_i)	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
Day/Good	900	824	5776	7.01
Day/Defective	130	206	5776	28.01
Evening/Good	700	696	16	0.02
Evening/Defective	170	174	16	0.09
Night/Good	400	480	6400	13.33
Night/Defective	200	120	6400	53.33
				$\chi^2 = 101.79$

Degrees of freedom = $(r - 1)(c - 1) = (3 - 1)(2 - 1) = 2$

Tabulated $\chi^2 (0.05, 2) = 5.991$

Decision Rule:

Since the calculated value of χ^2 is greater than the tabulated value of χ^2 , we may reject the null hypothesis and accept the alternative hypothesis, at 5% level of significance.

Conclusion:

There seems to be some association between the shifts and the quality of parts produced.

Example 10.12: 1000 students at college level were graded according to their IQ and economic conditions of their home. Use χ^2 tests to find out, whether there is any association between economic condition at home and IQ. (MBA, Osmania, 1996; MBA, Kumaun Univ., 1999)

Economic Condition ↓	I.Q.		Total
	High	Low	
Rich	460	140	600
Poor	240	160	400
Total	700	300	1000

Solution:

The null hypothesis:

H_0 : There is no association between economic condition at home and I.Q of students.

The alternative hypothesis

H_1 : There is association between economic condition at home and I.Q of students.

The observed frequencies are:

	I.Q.		
Economic Condition	High	Low	Total
Rich	460	140	600
Poor	240	160	400
Total	700	300	1000

The expected frequencies are:

	I.Q.		
Economic Condition	High	Low	Total
Rich	420	180	600
Poor	280	120	400
	700	300	1000

The test statistic

$$\begin{aligned}\chi^2 &= \frac{(460 - 420)^2}{420} + \frac{(140 - 180)^2}{180} + \frac{(240 - 280)^2}{280} + \frac{(160 - 120)^2}{120} \\ &= 3.81 + 8.89 + 5.71 + 13.33 \\ &= 31.74\end{aligned}$$

$$\text{Tabulated } \chi^2 (1, 0.05) = 3.84$$

Decision:

The null hypothesis is rejected, at 5% level of significance.

Conclusion:

There seems to be association between the economic condition at home and IQ of students.

10.2.3 Yates Correction for Continuity

The chi - square distribution is a continuous distribution. However, if any of the expected frequencies is less than 5, it fails to maintain continuity.

Particularly, in case of a 2×2 contingency table, the number of degrees of freedom is

$$(2 - 1) (2 - 1) = 1$$

In such a table, if any of the cell frequencies is less than 5, we pool it with another cell frequency. This pooling results in a loss of 1 d.f and thus, the total d.f. = 0, which renders the test invalid.

F. Yates (1934) gave a correction for this unique problem, and this correction is known as “Yates correction for continuity.”

This method consists in adding $\frac{1}{2}$ to the cell frequency less than 5 and then adjusting the remaining frequencies such that the marginal totals remain the same.

For a 2×2 contingency table of the form:

a	b
c	d

The chi - square statistic is:

$$\chi^2 = \frac{N(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)}$$

If suppose $a < 5$, then we add $\frac{1}{2}$ to a, subtract if from b & c and add $\frac{1}{2}$ to d, such that the marginal totals remain unchanged. Thus the 2×2 contingency table becomes:

$a + \frac{1}{2}$	$b - \frac{1}{2}$
$c - \frac{1}{2}$	$d + \frac{1}{2}$

The chi-square statistic, thus calculated is:

$$\text{corrected } \chi^2 = \frac{N \left[\left| ad - bc \right| - \frac{N}{2} \right]^2}{(a + c)(b + d)(a + b)(c + d)} \sim \chi_{(1)}^2$$

Example 10.13: The following data is related to a random sample of people attending a review of a new movie. A 2×2 contingency table is given as:

	Age \leq 30	Age $>$ 30
Liked the movie	32	08
Did not like the movie	04	06

Test whether the movie has equal appeal among people who are more than 30 years old and people who are less than 30 years old.

Solution:**Hypothesis**

H_0 : The movie has equal appeal among the two age groups.

H_1 : The movie does not appeal equally among the two age groups.

The Test Statistic

Since one of the cell frequencies is less than 5, we apply the corrected chi-square formula obtained by applying the Yates correction for continuity.

$$\begin{aligned} \text{corrected } \chi^2 &= \frac{N \left[|ad - bc| - \frac{N}{2} \right]^2}{(a+b)(a+c)(b+d)(c+d)} \\ &= \frac{50 \left[|32 \times 06 - 8 \times 4| - \frac{50}{2} \right]^2}{(32+8)(4+6)(32+4)(8+6)} \\ &= \frac{50 \left[|192 - 32| - 25 \right]^2}{40 \times 10 \times 36 \times 14} \\ &= \frac{50 \times 18225}{201600} \\ &= 4.52 \end{aligned}$$

Tabulated $\chi^2 (1, 0.05) = 3.841$

Decision Rule

Since calculated $\chi^2 >$ Tabulated $\chi^2 (1, 0.05)$, the null hypothesis may be rejected at 5% level of significance.

Conclusion

The movie has equal appeal among both the age groups.

10.2.4 Chi – Square test of Goodness of Fit

One more application of the chi – square statistic is in testing the goodness of fit of a certain hypothesized distribution. It determines if there is any significant difference between an observed frequency distribution and an expected frequency distribution.

The observed frequencies are calculated from the sample and the expected frequencies are calculated according to the hypothesized probability distribution.

In the chapter on probability distributions, we discussed fitting of the various distributions viz Binomial, Poisson and Normal Distributions.

Chi - square test of goodness of fit can be used to test how good a fit a particular distribution is for a given data set.

The steps of testing 'goodness-of-fit' is same as the testing for independence of attributes, except for three differences:

(i) Firstly, the null and the alternative hypothesis are as follows.

H_0 : The given distribution is a good fit to the data or the hypothesized probability distribution for a population provides a good fit.

H_1 : The hypothesized probability distribution for a population do not provide a good fit.

(ii) Secondly, the expected frequencies are given by

$$E(x) = Np(x) \text{ where } N = \sum_{i=1}^n f(x_i) = \text{sum of the frequencies}$$

$P(x)$ are the probabilities calculated according to the hypothesized population distribution.

(iii) The Chi-square statistic has $(n - 1)$ degrees of freedom.

Important Remark

If any of the expected frequencies are less than 5, then we have to resort to the 'pooling method'. This means we pool the observation which is less than 5 with frequencies before or after it such that after pooling the value exceeds 5.

The degrees of freedom would have to be adjusted accordingly.

For example, if $n = 10$, and two observations are pooled then 1 degree of freedom is subtracted from $n - 1$ and total d.f = $n - 1 - 1$

In case 3 observations are pooled

$$\text{d.f.} = n - 1 - 2$$

and so on.

We discuss this test with the help of an example first.

For example, suppose we have to test if a die is fair or not. The null hypothesis may be stated as

H_0 : The die is fair

$$\text{i.e. } p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}$$

where p_i = probability of getting a i

This is our hypothesized or expected frequency distribution when the die is fair.

To test this hypothesis, suppose we roll the die 150 times. The experiment results in the following observed frequencies.

Table 10.7
Observed Frequencies

i	1	2	3	4	5	6
Observed frequencies	20	30	25	20	40	15

In case our hypothesis is true we would expect each face to show up 25 times.

Thus, expected frequencies are:

Table 10.8
Expected Frequencies

i	1	2	3	4	5	6
Expected frequencies	25	25	25	25	25	25

Thus, the χ^2 test statistic is:

$$\begin{aligned}\chi^2 &= \frac{(20 - 25)^2}{25} + \frac{(30 - 25)^2}{25} + \frac{(25 - 25)^2}{25} + \frac{(20 - 25)^2}{25} + \frac{(40 - 25)^2}{25} + \frac{(15 - 25)^2}{25} \\ &= 1 + 1 + 0 + 1 + 9 + 4 \\ &= 16\end{aligned}$$

The number of degrees of freedom is $(6 - 1) = 5$. This is because there are six cells, but the number in the last cell can be determined automatically due to the constraint that the total should

add up to 120, because of the constraint $\sum_{i=1}^n O_i = \sum_{i=1}^n E_i$

Tabulated value of χ^2 at 5 degrees of freedom & 5% level of significance = 11.070.

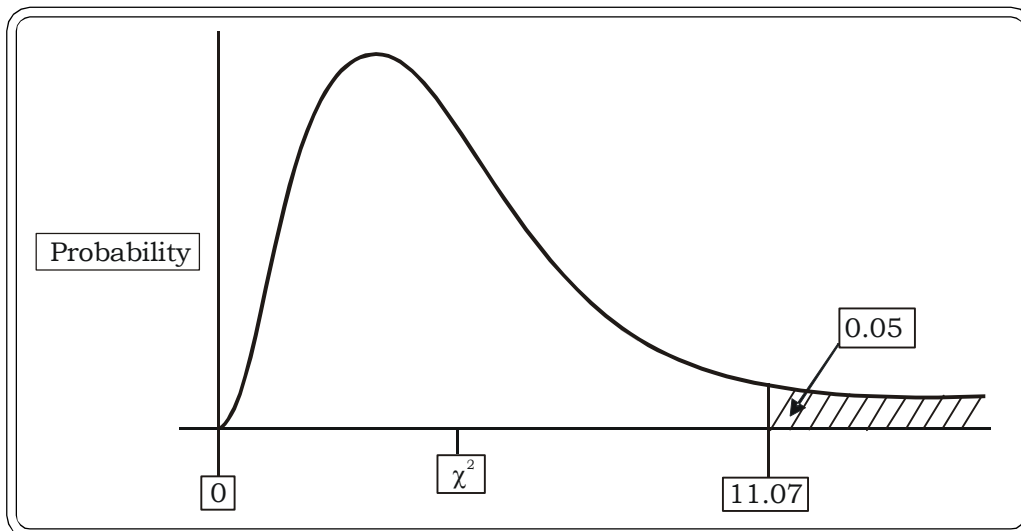


Figure 10.8
Graph of a χ^2 - distribution

Decision:

Thus, we will reject the null hypothesis and accept the alternative hypothesis at 5% level of significance.

Conclusion:

At 5% level of significance, there is not enough evidence to believe that the die is fair.

Example 10.14: A physicist theorizes that the probabilities that a radioactive substance emits 0, 1, 2, 3, or 4 particles of a certain kind during a one – hour period are, respectively, 0.05, 0.3, 0.3, 0.25 and 0.1. A record of 200 one – hour periods gave the following distribution of the number of particles emitted:

Number of Particles	0	1	2	3	4
Number of Items	16	68	51	46	19

At 5% level of significance, test the null hypothesis that there is no significant departure from what the physicist theorizes.

Solution:**The null hypothesis**

H_0 : There is no significance departure from what the physicist theorizes and what the sample data indicates.

H_1 : There is significant difference between what the physicist theorizes and what the sample data indicates.

Expected Frequencies

The expected frequencies are calculated as follows:

Number of Particles	Probability $p(x)$	Expected Frequencies $[N p(x)]$ $N = 200$
0	0.05	$E_1 = 200 \times 0.05 = 10$
1	0.3	$E_2 = 200 \times 0.3 = 60$
2	0.3	$E_3 = 200 \times 0.3 = 60$
3	0.25	$E_4 = 200 \times 0.25 = 50$
4	0.1	$E_5 = 200 \times 0.1 = 20$

Since there are 5 classes, d. f. = $5 - 1 = 4$

The test statistic

We calculate the χ^2 statistic as follows:

Number of Particles	Observed Frequencies (O_i)	Expected Frequencies (E_i)	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
0	16	10	36	3.6
1	68	60	64	1.07
2	51	60	81	1.35
3	46	50	16	0.32
4	19	20	1	0.05
				$\chi^2 = 6.39$

Tabulated χ^2 (0.05,4) = 9.488.

Decision:

We may accept the null hypothesis. Since calculated χ^2 is smaller than tabulated χ^2 .

Conclusion:

There is no significant departure from what the physicist theorizes and the results of the sample data.

Example 10.15: A machine is supposed to mix three types of candy- caramel, butter- scotch, and coconut – chocolate- in the proportion 4:3:2. A sample of 270 pieces of candy was found to contain 135 caramel, 70 butter- scotch and 65 coconut – chocolate. At the 5% level of significance, test the null hypothesis that the machine is mixing the candy in the proportion 4:3:2.

Solution:

The null hypothesis

H_0 : The machine is mixing candy in the proportion 4:3:2

i.e. caramel: butterscotch: coconut chocolate = 4:3:2

H_1 : The machine is not mixing candy in the proportion 4:3:2 as it is supposed to

$$\text{Expected frequencies: Caramel} = \frac{4}{9} \times 270 = 120$$

$$\text{Butter scotch} = \frac{3}{9} \times 270 = 90$$

$$\text{Coconut Chocolate} = \frac{2}{9} \times 270 = 60$$

Observed Frequencies:

Caramel: 135

Butterscotch: 70

Coconut Chocolate: 65

The test statistic

$$\begin{aligned}\chi^2 &= \frac{(135 - 120)^2}{120} + \frac{(70 - 90)^2}{90} + \frac{(65 - 60)^2}{60} \\ &= 1.875 + 4.44 + 0.42 \\ &= 6.73\end{aligned}$$

Tabulated χ^2 (.05,2) = 5.991

Decision:

The null hypothesis is rejected.

Conclusion:

The machine is not mixing the three types candy in the proportion 4:3:2.

As mentioned before, the χ^2 test can also be applied to test if the theoretical frequencies of a distribution follow certain standard distributions, for example, Binomial, Poisson or Normal distribution. The following example illustrates this application.

Example 10.16: At a level of significance of 0.10, can we conclude that the following 400 observations follow a Poisson distribution. (MBA, IGNOU, June 2003).

No. of Hours	0	1	2	3	4	5
No. of Arrivals	20	57	98	85	78	62

Solution:

The observed frequencies are given as 20, 57, 98, 85, 78 and 62.

We need to find the expected frequencies. In this case the expected frequencies are given by

$$\begin{aligned}E_i &= N p(i) \\ i &= 1, 2, 3, 4, 5\end{aligned}$$

where $p(i)$ = Poisson probability of arrivals being in the i^{th} hour.

$$= \frac{\lambda^i e^{-\lambda}}{i!}$$

N = Total no. of observations or total of the frequencies.

where λ is the mean no. of arrivals and is calculated as $\frac{\sum xf}{\sum f}$

The null hypothesis

H_0 : The given data follows a Poisson distribution.

The alternative hypothesis

H_1 : The given data does not follow a Poisson distribution.

We now calculate the mean of the observed data:

No. of Hours (x)	No. of arrivals (f)	fx
0	20	0
1	57	57
2	98	196
3	85	255
4	78	312
5	62	310
		$\sum fx = 1130$

$$\text{Thus } \bar{x} = \frac{1130}{400} = 2.825$$

Thus we set $\lambda = 2.825 \cong 3$

And now we calculate the expected Poisson probabilities using the poisson distribution tables.

$$P(0) = 0.0498$$

$$P(1) = 0.1494$$

$$P(2) = 0.2240$$

$$P(3) = 0.2240$$

$$P(4) = 0.1680$$

$$P(5) = 0.1848$$

This is calculated by the formula $p(5) = 1 - p(0) - p(1) - p(2) - p(3) - p(4)$ since total probability is 1.

The expected frequencies are now obtained by multiplying these probabilities by 400.

$$E_1 = 20$$

$$E_2 = 59$$

$$E_3 = 90$$

$$E_4 = 90$$

$$E_5 = 67$$

$$E_6 = 74$$

The test statistic

$$\begin{aligned}\text{Thus } \chi^2 &= 0 + \frac{(57 - 59)^2}{59} + \frac{(98 - 90)^2}{90} + \frac{(85 - 90)^2}{90} + \frac{(78 - 67)^2}{67} + \frac{(62 - 74)^2}{74} \\ &= 0.07 + 0.71 + 0.28 + 1.81 + 1.95 \\ &= 4.82\end{aligned}$$

The d.f. = 6 - 1 - 1 = 4, 1 more degree of freedom is subtracted due to the fact that mean was calculated from the given data.

Tabulated χ^2 (0.10, 4) = 7.779

Decision:

We may accept the null hypothesis at 10% level of significance.

Conclusion:

The Poisson distribution provides a good fit to the data

Example 10.17: The following table gives the number of car accidents that occurred during various days of the week. Find whether the accidents are uniformly distributed over the week.

(M.Com, M.D. Univ., 1999, MBA, DU, 2001)

Day	Sun.	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
No. of accidents	14	16	8	12	11	9	14

Solution:**The null hypothesis**

H_0 : The accidents are uniformly distributed over the week

The alternative hypothesis

H_1 : The accidents are not uniformly distributed over the week

Calculations:

Day	Observed No. of accidents (O_i)	Expected No. of accidents (E_i)	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
Sunday	14	12	4	0.33
Monday	16	12	16	1.33
Tuesday	8	12	16	1.33
Wednesday	12	12	0	0
Thursday	11	12	1	0.08
Friday	9	12	9	0.75
Saturday	14	12	4	0.33
				$\chi^2 = 4.15$

Calculated $\chi^2 = 4.15$

Tabulated $\chi^2 (.05, 6) = 12.592$

Decision:

We may accept the null hypothesis at 5% level of significance.

Conclusion:

The accidents are uniformly distributed over the week.

Example 10.18: L. Chandra, salesman for D. Paper Company, has 5 accounts to visit per day. It is suggested that the variable sales by Mr. Chandra may be described by the binomial distribution, with the probability of selling each account being 0.3. Given the following observed distribution of Chandra's number of sales per day, can we conclude that the distribution does in fact follow the suggested distribution? Use the 0.05 level of significance. (MFC, Delhi Univ., 1997)

No. of sales per day	0	1	2	3	4	5
Frequency of no. of sales	20	65	42	14	6	3

Solution:

The null hypothesis

H_0 : The variable sales may be described by a binomial distribution.

The alternative hypothesis

H_1 : The variable sales do not follow a binomial distribution.

The expected frequencies are given by

$$E_i = N P(i) \quad i = 1, 2, 3, 4, 5, N = 150$$

where $P(i) =$ Binomial Probability of no. of sales being i .

$$= \binom{n}{i} p^i q^{n-i}$$

n – no. of trials

p – probability of sales

$$q = 1 - p$$

$P(i)$ can be directly calculated from binomial tables for different values of i .

Here $p = 0.3$

Thus, from binomial tables

$$P(0) = 0.168$$

$$P(1) = 0.360$$

$$P(2) = 0.309$$

$$P(3) = 0.132$$

$$P(4) = 0.028$$

$$P(5) = 0.003$$

The expected frequencies are obtained by multiplying the P (i) by 150 (N)

$$E_1 = 25$$

$$E_2 = 54$$

$$E_3 = 46$$

$$E_4 = 20$$

$$E_5 = 4$$

$$E_6 = 1$$

Since E_5 and E_6 are less than 5, we need to apply the pooling technique here, before calculating the χ^2 test statistic.

E_4 and E_5 are pooled. The observed frequencies corresponding to these frequencies are also pooled simultaneously.

$$\text{And d.f.} = 6 - 1 - 1 = 4$$

1 Degree of freedom are lost due to pooling of two cell frequencies.

This is shown in the following table:

No. of sales per day	Observed Frequencies (O_i)	Expected frequencies (E_i)	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	20	25	25	1
1	65	54	121	2.24
2	42	46	16	0.35
3	14	20	36	1.8
4	6	4		
	9	5	16	3.2
5	3	1		
				8.59

Thus, calculated $\chi^2 = 8.59$

Tabulated $\chi^2 = (0.05, 4) = 9.488$

We may accept the null hypothesis at 5% level of significance.

Conclusion:

L. Chandra's sales may be adequately described by a binomial distribution.

10.3 ONE - WAY ANALYSIS OF VARIANCE (ANOVA)



In chapter 7, we discussed tests for comparing two population means by using a z - test, if the population variance is unknown and a t test if the population variance is unknown and the sample size is small. These ideas can be generalized to give a test for testing equality of several means. Such a test is called Analysis of Variance or ANOVA. Thus, ANOVA is used to test if more than two populations have the same mean value. We illustrate ANOVA with the help of an example.

Suppose we wish to compare three different brands of tyres. The number of kilometers when 3 tyres of each brand were tested is given in the following table:

Table 10.9
Mileage by Three Different Brands of Tyres

Brand A	Brand B	Brand C
$x_{11} = 30,000$	$X_{21} = 33,000$	$X_{31} = 36,000$
$x_{12} = 35,000$	$X_{22} = 28,000$	$X_{32} = 30,000$
$x_{13} = 41,000$	$X_{23} = 38,000$	$X_{33} = 40,000$

From this table, it is apparent that there is some variation within each brand and some variation among the three brands with respect to their means. The question arises as to whether these differences are due to chance or because of the fact that the brands are basically different from each other.

Thus, if μ_1 = mean mileage of Brand A

μ_2 = mean mileage of Brand B

μ_3 = mean mileage of Brand C

Then, the null hypothesis is

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

i.e. there is no significant difference in the mean number of kilometers each brand can last.

The **alternative hypothesis**

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$$

i.e. there is significant difference in the performance of the three brands with respect to the mean number of kilometers they can run.

Let \bar{x}_1 = mean number of kilometers of Brand A during trials

\bar{x}_2 = mean number of kilometers of Brand B during trials

\bar{x}_3 = mean number of kilometers of Brand C during trials

Then $\bar{x}_1 = 35,333$; $\bar{x}_2 = 33,000$; $\bar{x}_3 = 35,333$

and $\bar{\bar{x}} = 34,555$ (grand mean)

In ANOVA, the total variability in the entire data set is divided into two kinds:

- (a) Variability within the brands &
- (b) Variability between the brands

Thus,

Total variation = Variation within the brands + Variation between the brands

First, the total sum of squares is expressed as

$$TSS = SSE + SSB \quad (1)$$

where TSS = Total sum of squares

$$\begin{aligned} &= (30,000 - 34,500)^2 + (35,000 - 34,500)^2 + (41,000 - 34,500)^2 \\ &+ (33,000 - 34,500)^2 + (28,000 - 34,500)^2 + (38,000 - 34,500)^2 \\ &+ (36,000 - 34,500)^2 + (30,000 - 34,500)^2 + (40,000 - 34,500)^2 \end{aligned}$$

SSE = Error sum of squares or within brands sum of squares

$$\begin{aligned} &= (30,000 - 35,333)^2 + (35,000 - 35,333)^2 + (41,000 - 35,333)^2 \\ &+ (33,000 - 33,000)^2 + (28,000 - 33,000)^2 + (38,000 - 33,000)^2 \\ &+ (36,000 - 35,333)^2 + (30,000 - 35,333)^2 + (40,000 - 35,333)^2 \end{aligned}$$

SSB = between sum of squares or sum of squares between the brands.

$$= 3 (35,333 - 34,555)^2 + (33,000 - 34,555)^2 + 3 (35,333 - 34,555)^2$$

In practice any two quantities in equation (1) are computed and the remaining quantity is calculated by subtraction.

In theory, the quantities in equation (1) can be expressed as

$$SSE = \sum_{i=1}^3 \sum_{j=1}^3 (x_{ij} - \bar{x}_i)^2 \quad \text{and}$$

$$SSB = n \sum_{i=1}^3 (\bar{x}_i - \bar{\bar{x}})^2$$

where x_{ij} - j^{th} observation in the i^{th} brand (or category)

\bar{x}_i - mean of the i^{th} brand (category)

$$i = 1, 2, 3$$

$$j = 1, 2, 3$$

The degrees of freedom are:

$$\text{For TSS} = \sum n_i - 1 = n - 1$$

$$\text{For SSB} = 3 - 1$$

Next, unbiased estimates of the variations are obtained.

$$\text{Unbiased estimator of Total Variation} = \frac{\sum_{i=1}^3 \sum_{j=1}^3 (x_{ij} - \bar{x})^2}{\left(\sum_{i=1}^3 n_i - 1 \right)}$$

$$\text{Unbiased estimator of Variation within the brands (MSE)} = \frac{\sum_{i=1}^3 \sum_{j=1}^3 (x_{ij} - \bar{x}_i)^2}{3(n-1)}$$

$$\text{Unbiased estimator of between within the brands (MSB)} = \frac{n \sum_{i=1}^3 (\bar{x}_i - \bar{x})^2}{(3-1)}$$

k being the number of brands being compared and k = 3 in this example.

Thus,

$$\begin{aligned} \text{TSS} &= 20.25 + 0.25 + 42.25 + 2.25 + 42.25 + 12.25 + 2.25 + 20.25 + 30.25 \\ &= 172.25 \text{ (x 1,000, 000)} \end{aligned}$$

$$\begin{aligned} \text{SSE} &= 28.09 + 0.09 + 22.09 + 0 + 25 + 25 + 0.49 + 28.09 + 22.09 \\ &= 161.33 \end{aligned}$$

$$\text{SSB} = 10.89$$

and d.f. for SSE = 6 and for SSB = 2

$$\text{MSE} = 26.89$$

$$\text{MSB} = 5.44$$

The test statistic

Since we are comparing two variances we use a F - statistic often referred to as the F- ratio and defined as

$$F = \frac{\text{MSB}}{\text{MSE}} = 0.2025, \text{ which is less than 1. Therefore we invert it.}$$

$$F' = \frac{1}{F}$$

This statistic follows a F distribution with 3 (n - 1) and (n - 1) degrees of freedom.

$$\text{Thus } F' = 4.9383$$

This is the calculated value of F.

If the calculated value of F exceeds the table value of the F - distribution for a given level of significance we may reject the null hypothesis and accept it otherwise.

In this case

Calculated $F = 4.9383$

and table value of F at degrees of freedom (6, 2) and level of significance 0.05 is 19.33.

All this information is summarized in the Analysis of Variance table below:

Table 10.10
ANOVA Table

Sources of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of squares	F-ratio
Between the Groups	$k - 1$	SSB	MSB	$\frac{MSB}{MSE}$
Within the Groups	$3(n - 1)$	SSE	MSE	
Total	$\sum n_i - 1$	SST		

In this example the ANOVA table with computed values are

Table 10.11
ANOVA Table

Sources of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of squares	F - ratio
Between	2	21.31	10.65	0.42
Within	6	150.94	25.16	
Total	8	172.25		

Decision:

We may accept the null hypothesis at 5% level of significance.

Conclusion:

Thus, there is no significant difference in the performance of the three brands of tyres.

Generalization

In general, suppose the data consists of k groups, each group containing n_1, n_2, \dots, n_k observation as follows:

Groups	1	2	k
	x_{11}	x_{21}	x_{k1}
	x_{12}	x_{22}	x_{k2}
	\vdots	\vdots	\vdots
	x_{1n_1}	x_{2n_2}	x_{kn_k}

Table 10.12
Layout of Observations

Groups	1	2			k
	x_{11}	x_{12}	x_{k1}
	x_{21}	x_{22}	x_{k2}
	
	x_{1n_1}	x_{2n_2}	x_{kn_k}

Sample sizes: n_1 n_2 n_k

Sample means: \bar{x}_1 \bar{x}_2 \bar{x}_k

The total number of observations is $\sum_{i=1}^k n_i = N$

$$TSS = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{\bar{x}})^2, \quad x_{ij} = j^{\text{th}} \text{ observation in the } i^{\text{th}} \text{ group}$$

$\bar{\bar{x}}$ = grand mean

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

$$SSB = \sum_{i=1}^k n_i (\bar{x}_i - \bar{\bar{x}})^2$$

$$MSB = \frac{\sum_{i=1}^k n_i (\bar{x}_i - \bar{\bar{x}})^2}{k - 1}$$

$$MSE = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{N - k}$$

Thus, the ANOVA table is:

Table 10.13
ANOVA Table

Sources of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of squares	F - ratio
Between	$k - 1$	SSB	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSE}$
Within	$N - k$	SSE	$F = \frac{MSB}{MSE}$	
Total	$N - 1$	TSS		

The decision rule remains the same.

10.4 ASSUMPTIONS OF ANOVA

Analysis of variance rests on three assumptions:

1. Randomness

The samples drawn must be random samples

2. Normality Assumption

The population from which the examples are drawn must be normally distributed.

3. Homogeneity of Variance

Third assumption is that the populations have equal variances.

10.5 SIMPLE STEPS FOR ANOVA CALCULATIONS

The following steps may be followed for easy calculation of the terms of the ANOVA table.

Step 1: Calculate what is called the Correction Factor (CF) by using the formula

$$CF = \frac{(\text{Grand Total})^2}{N} = \frac{(x_{..})^2}{N}, \text{ where } x_{..} = \text{Sum of all the observations}$$

$$\text{Step 2: } SSB = \frac{x_{1.}^2}{n_1} + \frac{x_{2.}^2}{n_2} + \frac{x_{3.}^2}{n_3} + \frac{x_{4.}^2}{n_4} + \dots + \frac{x_{k.}^2}{n_k} - CF$$

$x_{i.}$ - total of the i^{th} group; $i = 1, 2, \dots, k$.

$$\text{Step 3: } TSS = \sum_{i=1}^k \sum_{j=1}^{n_j} x_{ij}^2 - CF, \text{ } x_{ij} - j^{\text{th}} \text{ observation in the } i^{\text{th}} \text{ group}$$

Step 4: $SSE = TSS - SSB$

The remaining quantities are calculated as usual.

Example 10.19: To assess the significance of possible variation in performance among four UPSC training centers in a city, a common test was given to a number of students taken at random and the results are given below. Apply ANOVA for testing the variation in performance among the four training centers.

A	B	C	D
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15

Solution:

The null and the alternative hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

i.e. the average scores of the students in all the four centers are same.

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$$

i.e. the average scores of the students in all the four centers are different.

Calculations:

We may use the simple steps to complete the ANOVA Table:

	A	B	C	D	
	8	12	18	13	
	10	11	12	9	
	12	9	16	12	
	8	14	6	16	
	7	4	8	15	
x_i	45	50	60	65	220
\bar{x}_i	9	10	12	13	

$x..$ = Sum of all the observations = 220

Here k = no of treatments. In this example, no of training centres
= 4

$$n_1 = n_2 = n_3 = n_4 = 5$$

$$N = \sum_{i=1}^4 n_i = 20$$

From the table $x_1 = 45$

$$x_2 = 50$$

$$x_3 = 60$$

$$x_4 = 65$$

Step 1: We first calculated the correction factor

$$\text{C.F.} = \frac{x_{..}^2}{N} = \frac{220^2}{20} = 2420$$

Step 2: We now calculate SSB using the formula.

$$\begin{aligned} \text{SSB} &= \left(\frac{x_1^2}{n_1} + \frac{x_2^2}{n_2} + \frac{x_3^2}{n_3} + \frac{x_4^2}{n_4} \right) - \text{C.F.} \\ &= \frac{1}{5} (45^2 + 50^2 + 60^2 + 65^2) - 2420 \\ &= \frac{1}{5} (2025 + 2500 + 3600 + 4225) - 2420 \\ &= 2470 - 2420 \\ &= 50 \end{aligned}$$

Thus SSB = 50

Step 3: We next calculate SST

$$\begin{aligned} \text{TSS} &= \sum_{i=1}^4 \sum_{j=1}^{n_i} x_{ij}^2 - \text{C.F.} \\ &= 2678 - 2420 \\ &= 258 \end{aligned}$$

Step 4: By subtraction

$$\begin{aligned} \text{SSE} &= \text{TSS} - \text{SSB} \\ &= 258 - 50 \\ &= 208 \end{aligned}$$

Step 5: Degrees of Freedom

$$\text{D.f. for SSB} = k - 1 = 4 - 1 = 3$$

$$\text{D.f. for TSS} = N - 1 = 20 - 1 = 19$$

$$\text{D.f. for SSE} = N - k = 20 - 4 = 16$$

Step 6: Calculation of MSS

$$MSB = \frac{SSB}{3} = \frac{50}{3} = 16.67$$

$$MSE = \frac{SSE}{16} = \frac{208}{16} = 13$$

Step 7: The F - Ratio

$$F = \frac{MSB}{MSE} = 1.28$$

All the calculations are now summarized in the following ANOVA Table.

The ANOVA Table

Sources of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of squares	F - ratio
Between Groups	3	50	16.67	1.28
Within Groups	16	208	13	
Total	19	258		

Tabulated value of F (.05, 3, 16) = 3.24

Decision:

The null hypothesis may be accepted at 5% level of significance.

Conclusion:

The average scores of the students in all the four centers are same. Thus, there is no significant variation in performance of the students among the four UPSC training centers in the city.

Example 10.20: A bank manager wanted to compare the time taken by four tellers to attend to customers. The following data was compiled regarding the amount of time (in minutes) that they spent serving each customer. The sample values have been recorded randomly.

	Teller 1	Teller 2	Teller 3	Teller 4
1	1	8	9	5
2	2	5	3	2
3	8	1	6	2
4	7	9		4

Is there any significant difference in the mean client serving time of the four tellers (use 0.05 level of significance)?

Solution:

Let

 μ_1 = mean client servicing time of teller 1 μ_2 = mean client servicing time of teller 2 μ_3 = mean client servicing time of teller 3 μ_4 = mean client servicing time of teller 4**The null hypothesis** $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ i.e. the mean client serving time of all the four tellers are same**The alternative hypothesis** $H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$ i.e. the mean client serving time of all four tellers different

Calculations for ANOVA

Teller 1	Teller 2	Teller 3	Teller 4	
1	8	9	5	
2	5	3	2	
8	1	6	2	
7	9		4	
$n_1 = 4$	$n_2 = 4$	$n_3 = 3$	$n_4 = 4$	
$\bar{x}_1 = 4.5$	$\bar{x}_2 = 5.75$	$\bar{x}_3 = 4.5$	$\bar{x}_4 = 3.25$	
$x_{1.} = 18$	$x_{2.} = 23$	$x_{3.} = 18$	$x_{4.} = 13$	$x_{..} = 72$

Grand mean = 4.5

 $N = 15, k = 4$ $x_{i.}$ = Total time taken by the i^{th} teller, $i = 1, 2, 3, 4$.

Simple Steps for ANOVA

The following steps are now followed:

Step 1: We first calculate the Correction Factor (CF) by using the formula:

$$CF = \frac{(\text{Grand Total})^2}{N} = \frac{(x_{..})^2}{N} = 345.6$$

$$\begin{aligned}
 \text{Step 2: SSB} &= \frac{x_{1.}^2}{n_1} + \frac{x_{2.}^2}{n_2} + \frac{x_{3.}^2}{n_3} + \frac{x_{4.}^2}{n_4} - CF \\
 &= 81 + 132.25 + 108 + 42.25 - CF \\
 &= 363.5 - 345.6 \\
 &= 17.9
 \end{aligned}$$

$$\begin{aligned}
 \text{Step 3: TSS} &= \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}^2 - CF \\
 &= 464 - 345.6 \\
 &= 118.4
 \end{aligned}$$

$$\begin{aligned}
 \text{Step 4: SSE} &= \text{TSS} - \text{SSB} \\
 &= 100.5
 \end{aligned}$$

The ANOVA table is as follows:

ANOVA Table

Sources of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of squares	F - ratio
Between the tellers	3	17.9	5.97	0.605
Error (within the tellers)	11	100.5	9.13	
Total	14	118.4		

Tabulated value of F (.05, 3,11) = 2.66

Decision:

The null hypothesis may be accepted at 5% level of significance.

Conclusion:

There is no significant difference in the mean client servicing time of the four tellers.

Example 10.21: An organization provides training to company employees. They follow three training methods. They want to compare the three training methods to check which one leads to greater productivity among the employees. The following numbers are the productivity measures for the individuals trained by each method.

Method A	25	30	28	32	33
Method B	30	28	21	25	27
Method C	32	35	28	32	30

Use ANOVA techniques to determine if the three training methods are producing different results.

Solution:

H_0 : The three training methods do not produce different results.

H_1 : The three training methods lead to different levels of productivity.

Calculations

Step 1: Grand Total = $x_{..}$ = 436

$$CF = \frac{(436)^2}{15} = 12673.07$$

$N = 15, k = 3$

Method A	Method B	Method C
25	30	32
30	28	35
28	21	28
32	25	32
33	27	30
$x_{1.} = 148$	$x_{2.} = 131$	$x_{3.} = 157$
$\bar{x}_1 = 29.6$	$\bar{x}_2 = 26.2$	$\bar{x}_3 = 31.4$

Step 2:

$$SSB = \left(\frac{x_{1.}^2}{n_1} + \frac{x_{2.}^2}{n_2} + \frac{x_{3.}^2}{n_3} \right) - CF$$

$$= (4380.8 + 3432.2 + 4929.8) - CF$$

$$= 12742.8 - 12673.07$$

$$= 69.73$$

Step 3:

$$TSS = 12858 - 12673.07$$

$$= 184.93$$

Step 4:

$$SSE = SST - SSB$$

$$= 115.2$$

ANOVA Table

Sources of Variation	D.F.	S.S.	M.S.S.	F - ratio
Between the methods	2	69.73	34.87	3.63
Error	12	115.2	9.6	
Total	14	184.93		

Tabulated F (2,12, 0.05) = 3.89

Decision:

We may accept the null hypothesis, at 5% level of significance.

Conclusion:

At 5% level of significance, we may conclude that the three training methods do not lead to different productivity levels among the employees. i.e. the three training methods are equally effective.

Example 10.22: The following are the number of mistakes made in 5 successive days by 4 technicians working for a photographic laboratory. Test at a level of significance $\alpha = 0.01$, whether the differences among the four sample means can be attributed to chance. (MBA, Anna Univ., 2003)

Mistakes	Technician I	Technician II	Technician III	Technician IV
Day 1	6	14	10	9
Day 2	14	9	12	12
Day 3	10	12	7	8
Day 4	8	10	15	10
Day 5	11	14	11	11

Solution:**Null and alternative hypothesis**

H_0 : There is no significant difference among the technicians in terms of their mean number of mistakes.

H_1 : There is significant difference among the technician performance.

Calculations:

k = number of technicians = 4

N = Total number of observations = 20

x_i = Total mistakes made by i^{th} technician $i = 1, 2, 3, 4$

$x_{1.}$ = 49

$x_{2.}$ = 59

$x_{3.}$ = 55

$x_{4.}$ = 50

$x_{..}$ = Total number of mistakes made by all the technicians
= 213

Step 1: The correction factor (CF)

$$CF = \frac{(\text{Grand Total})^2}{N} = \frac{(213)^2}{20} = 2268.45$$

Step 2: SSB

$$\begin{aligned} SSB &= \left(\frac{49^2}{5} + \frac{59^2}{5} + \frac{55^2}{5} + \frac{50^2}{5} \right) - CF \\ &= (480.2 + 696.2 + 605 + 500) - 2268.45 \\ &= 2281.4 - 2268.45 \\ &= 12.95 \end{aligned}$$

Step 3: TSS

$$\begin{aligned} TSS &= 2383 - 2268.45 \\ &= 114.55 \end{aligned}$$

Step 4: SSE = TSS - SSB

$$\begin{aligned} &= 114.5 - 12.95 \\ &= 101.6 \end{aligned}$$

ANOVA Table

Sources of Variation	D.F.	S.S.	M.S.S.	F - ratio
Between	3	12.95	4.32	
Error	16	101.6	6.32	0.68
Total	19	114.5		

Since the F value is less than 1, we use $F' = \frac{1}{F} = 1.46$

Tabulated $F(0.01, 16, 3) \cong 26.87$

Decision:

We may accept the null hypothesis at 1% level of significance.

Conclusion:

At 1% level of significance, we may conclude that there is no significant difference between the performances of the four technicians.

Example 10.23: During the last week, there were 14 sales calls. A made 5 calls, B made 4 calls and C made 5 calls. Following are the weekly sales (in 000's Rs.) record of the three salesmen:

	Salesmen		
	A	B	C
	3	6	7
Calls	4	3	3
	3	3	4
	5	4	6
	0	-	5

With the help of analysis of variance, test the selling ability of the three salesmen. (MBA, Delhi Univ., 1989)

Solution:

Null hypothesis and alternative hypothesis

H_0 : There is no significant difference in the selling ability of the three salesmen.

H_1 : There is significant difference in the selling ability of the three salesmen.

Calculations:

Step 1: The correction factor

$$CF = \frac{(\text{Grand Total})^2}{N} = 14 = 224, N = 14$$

Step 2: SSB. Here $n_1 = 5, n_2 = 4, n_3 = 5$

$$\begin{aligned} SSB &= \left(\frac{(15)^2}{5} + \frac{(16)^2}{4} + \frac{(25)^2}{5} \right) - CF \\ &= 234 - 224 \\ &= 10 \end{aligned}$$

Step 3: TSS

$$\begin{aligned} TSS &= 264 - 224 \\ &= 40 \end{aligned}$$

Step 4: SSE

$$\begin{aligned} SSE &= TSS - SSB \\ &= 30 \end{aligned}$$

The ANOVA Table

Sources of Variation	D.F.	S.S.	M.S.S.	F - ratio
Between the salesmen	2	10	5	1.83
Error	11	30	2.73	
Total	13	40		

Tabulated value of F (2, 11, 0.05) = 3.98

Decision:

We may accept the null hypothesis at 5% level of significance.

Conclusion:

At 5% level of significance, there is no significant difference in the selling ability of the three salesmen.

10.6 CASELETS

Caselet 1: A company sells insurance policies to its customers. For selling its policies, the company appoints recruitment agents who can there-by sell the policies. The requirements to qualify as insurance advisor for the insurance company are

- (i) they should be graduate in any stream.
- (ii) they should be above 25 years in age.
- (iii) they should be living in the city for more than 5 years

Further, after these initial conditions are fulfilled, they also have to clear an examination by the board, which governs all insurance related matters. The company is using the methods to recruit these agents viz. cold calling, reference appointment and through stalls at various market places frequented by crowds. The senior HR manager over-seeing the recruitment process wants to evaluate whether all the three approaches are equally effective or not. He gathers the following data:

Recruitment Approach

	Cold Calling	Reference Calling	Stall/Canopy
Appointed	6	13	20
Not Appointed	34	30	80
Total	40	43	100

The manager also wants to evaluate whether the recruitment approach and the number of people appointed show any kind of association.

How can you help the manager answer these questions?

Caselet 2: A fast food chain promises to deliver customer orders within 5 minutes of customers placing the order. The chain has opened its outlets at 4 major cities of India viz. Mumbai, Delhi, Chennai and Bangalore. The company has major concerns about deliveries being on time as promised. It also wants to evaluate its performance in the four cities where they have just started their business. The performance of the chains in the four cities is closely monitored for a month to identify potential areas that need improvement. Data taken randomly from the four outlets over a period of one month gave the following service time/minutes:

Mumbai	Delhi	Chennai	Banglore
2.5	3.5	4.2	3.9
3.0	3.8	3.9	4.0
3.2	4.5	3.6	4.1
4.0	4.2	2.7	4.9
4.5	4.0	3.5	5.0
3.8	4.9	4.5	2.8
3.6	5.0	4.3	2.2
3.2	3.0		

What conclusions can you draw about the relative performance of the chain across the four cities?

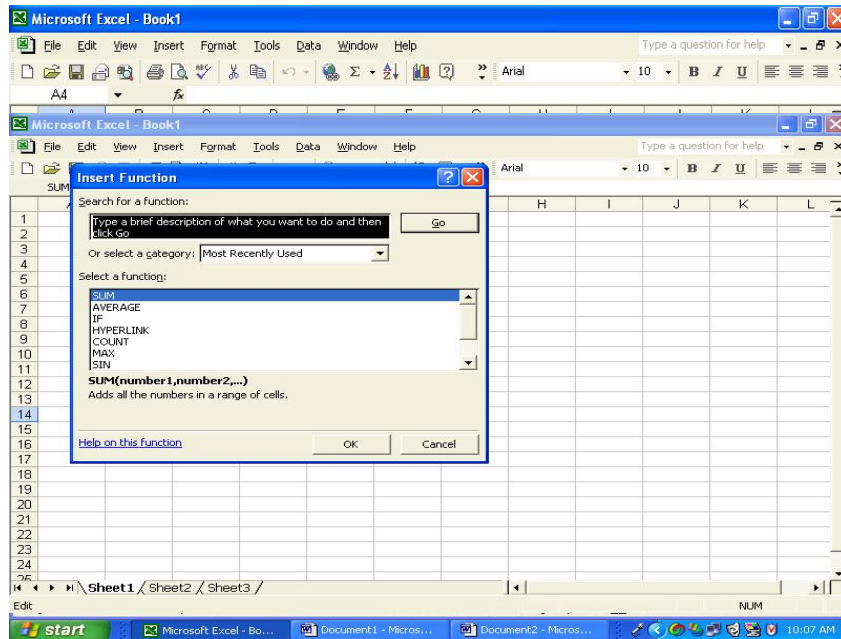
10.7 EXCEL GUIDE

Excel guide for χ^2 – Test for independence of two attributes

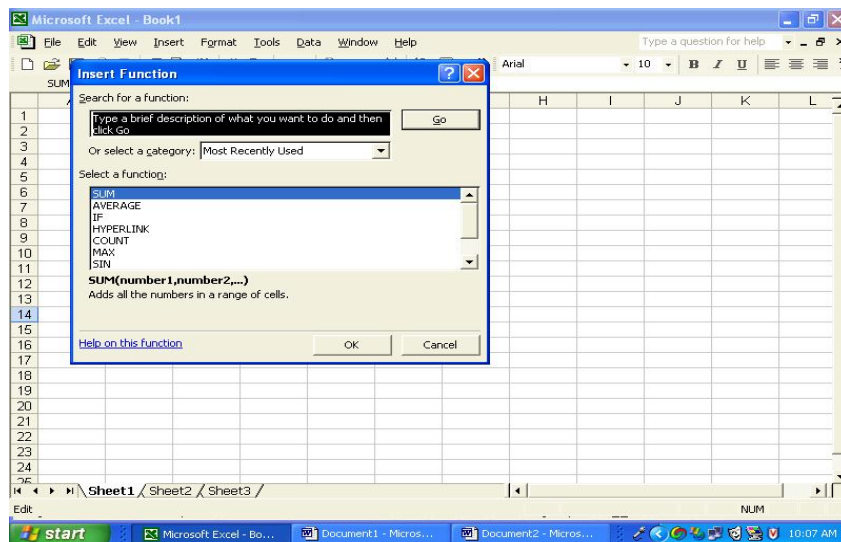
Step 1: Enter the observed frequencies from cell A6: A14 and the expected frequencies from B6: B14 as shown below in a worksheet.

The screenshot shows an Excel spreadsheet with a 2x8 contingency table. The first column (A6:A14) contains observed frequencies: 40, 75, 65, 30, 45, 75, 40, 100, 190. The second column (B6:B14) contains expected frequencies: 30, 60, 90, 25, 50, 75, 55, 110, 165. A dialog box titled "Function Arguments" for the CHITEST function is open, showing the actual range as A6:A14 and the expected range as B6:B14. The dialog box displays the formula result as 0.002028427. The dialog box also includes a description of the function: "Returns the test for independence: the value from the chi-squared distribution for the statistic and the appropriate degrees of freedom." and "Expected_range is the range of data that contains the ratio of the product of row totals and column totals to the grand total."

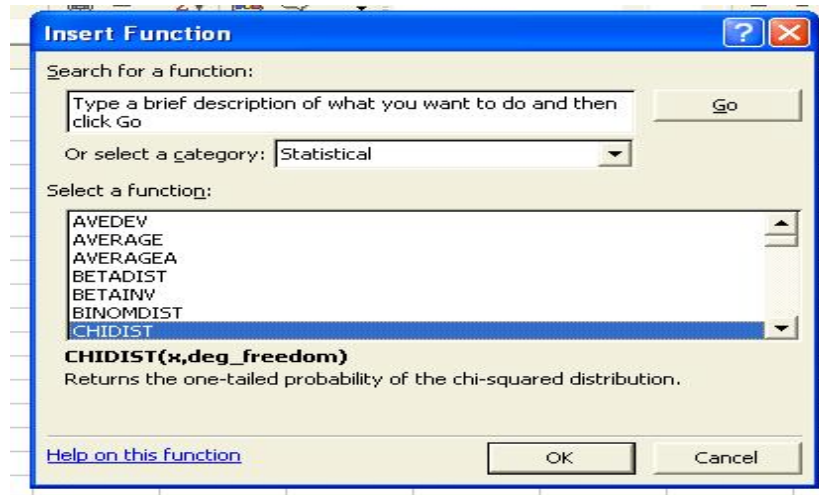
Step 2: Choose any cell in the excel worksheet to display the result. Here A4 has been chosen.



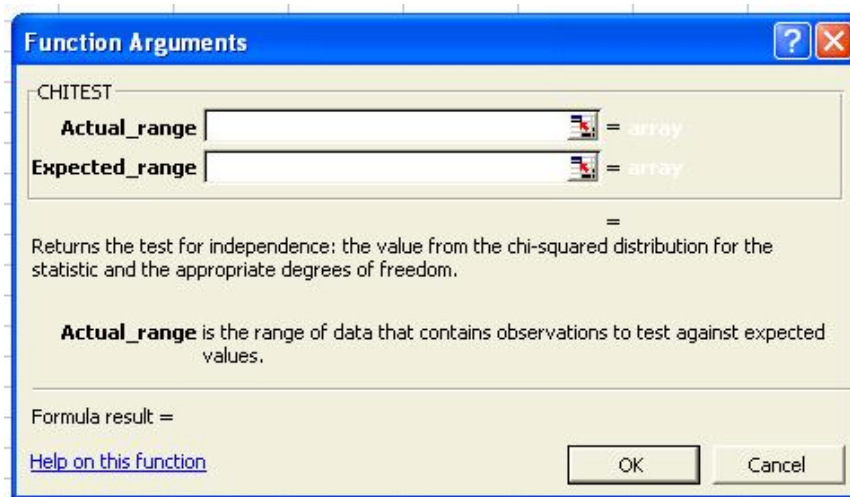
Step 3: Go to the formula bar or the INSERT FUNCTION option



Step 4: Select the category STATISTICAL and chose CHIDIST and click OK



Step 5: The following screen will be displayed. In Actual Range we type A6:A14 the observed frequencies and in expected range type B6:B14. Click OK.

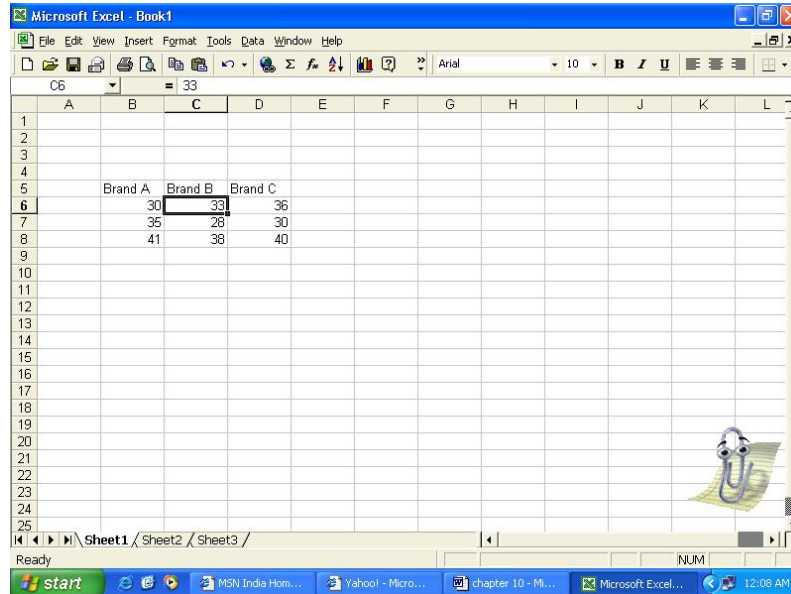


Step 6: The final result will be displayed in A4.

	A	B	C
1			
2			
3			
4	0.002028		
5	O	E	
6	40	30	
7	75	60	
8	65	90	
9	30	25	
10	45	50	
11	75	75	
12	40	55	
13	100	110	
14	190	165	
15			
16			

Excel Guide for One Way ANOVA

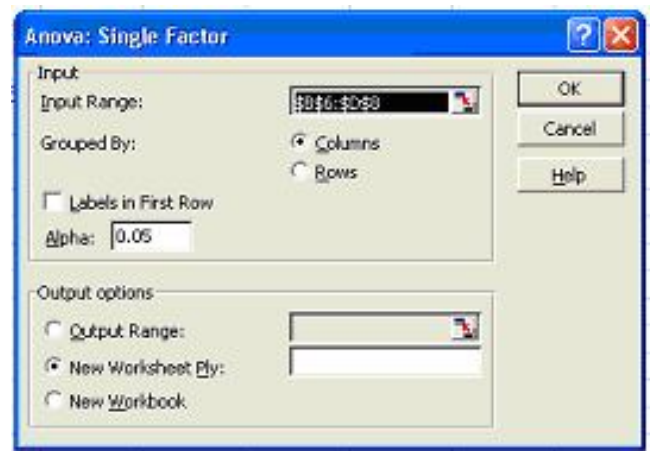
Step 1: Enter data in Excel spreadsheet as follows. Here we test for difference among three brands: Brand A, Brand B, Brand C.



Step 2: Go to TOOLS and select DATA ANALYSIS. Chose ANOVA:SINGLE FACTOR. Click OK.



Step 3: In input range type B6: D8 the entire data. In the option grouped by chose columns if data is in columns and rows if data has been entered in rows. Since the data here is in columns we chose columns. In output options chose NEW WORKSHEET PLY for the result to be displayed in a new worksheet. Click OK.



Step 4: The result is displayed in worksheet no. 5 as shown below:

The screenshot shows an Excel spreadsheet with the following data:

Groups	Count	Sum	Average	Variance
Column 1	3	106	35.33333	30.33333
Column 2	3	99	33	25
Column 3	3	106	35.33333	25.33333

Source of Variance	SS	df	MS	F	P-value	F crit
Between Groups	10.88889	2	5.444444	0.202479	0.822062	5.143249
Within Groups	161.3333	6	26.88889			
Total	172.2222	8				

10.8 EXERCISES

- 10.1 What are the uses of a chi – square test?
- 10.2 Why is a chi – square test regarded as a non-parametric test?
- 10.3 Give examples of situations, which require application of chi – square test
 - (1) of goodness of fit.
 - (2) of independence
 - (3) of equality of proportion.
- 10.4 State the assumptions to be made in the one-way analysis of variance.
- 10.5 Give examples of situations where you can apply Analysis of Variance.
- 10.6 In an ANOVA problem with five categories, the $SSB = 10.12$, $SST = 50.13$. The total number of observations is 50. Find the F ratio and also give the ANOVA table.
- 10.7 To test the effectiveness of a drug, it was tested on 200 people with the disease. 100 of them were given the drug and the result was noted and 100 were not given the drug and were monitored to check if they recovered from the disease or not. The following results were obtained.

	Drug	No Drug
Recovered	80	50
Not Recovered	20	50
	100	100

- 10.8 The inventory of a particular component in a factory seemed to vary a lot from day to day. The management in a bid to study this problem monitored it over a period of a week and got the following results.

Day	1	2	3	4	5	6	7
No. of Components	2000	2200	2100	3000	2500	2800	2900

Does the requirement of the component depend on the day of the week?

- 10.9 A mobile phone company is launching a new model in the market. The marketing team wants to know if the model will be equally appealing across all age groups or not. They conduct a survey of 1000 people across different age groups to find out whether the model has uniform appeal across all ages. The following data is compiled.

Age Groups

Under	20	20-20	30-40	40-50	50 & above
Liked the model	150	100	95	80	90
Do not like the model	50	100	105	120	110
Total	200	200	200	200	200

Is there reason to believe that model appeal is independent of age group?

- 10.10 A production manager wants to test if the number of bad parts produced depends on the production shift. He compiles the following data related to parts and the shift from which they come from.

	No. of goods parts	No. of bad parts
Day Shift	70	20
Evening Shift	60	15
Night Shift	70	15
	200	50

At 5% level of significance, test if the number of bad parts produced depend on the production shift.

- 10.11 A company manufacturing water filters wants to find out if its 4 salesmen have the same ability to sell the product or their performances vary. Data on the number of water filters sold by the 4 salesmen over a period of 4 weeks on last month was gathered.

	Salesman 1	Salesman 2	Salesman 3	Salesman 4
1 st Week	40	28	35	50
2 nd Week	50	30	38	20
3 rd Week	30	32	40	30
4 th Week	35	37	42	35

Test if all the salesmen have equal sales potential.

- 10.12 A school has 4 sections in its 12th standard A, B, C, and D. Each section is taught by a different teacher. In the final examinations, the scores of the students are obtained as follows:

Section A	Section B	Section C	Section D
85	75	82	85
80	78	81	84
90	76	80	90
92	85	95	88
100	100	100	100

Find out if there is any relationship between the scores and sections (use 5% level of significance).

- 10.13 A study was conducted to determine if smokers suffering from asthma were more prone to attacks during winter. Of 100 smokers surveyed, 65 replied in the affirmative and of 100 non-smokers 80 said yes and the rest replied in the negative. At 5% level of significance is proportion of people prone to attacks same in the winter month.
- 10.14 To find the attitude of people on the issue of reservation for backward classes. 100 people were interviewed in Delhi, Mumbai, Kolkata & Chennai. The following data was obtained:

	Delhi	Mumbai	Kolkata	Chennai
In favor	25	20	30	35

At 5% level of significance, test the hypothesis test that the attitude of people on reservation is same in all the four cities.

- 10.15 An insecticide manufacturing company wants to test 4 different insecticides. Four jars containing 100 insects were sprayed with the four different insecticides. The number of insects killed were 50, 70, 80 and 90. Use 5% level of significance to test whether the four insecticides are different.
- 10.16 The following data is from a survey of 100 patients 60 of whom were administered certain drug and 40 of them were not given any drug.

	Drug	No Drug
Cured	45	28
Not Cured	15	12
	60	40

10.17 Complete the following ANOVA table.

Sources of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of squares	F - ratio
Between	3			
Error		35		
Total	29	57		

10.18 Complete the following ANOVA table.

Sources of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of squares	F - ratio
Between				
Error		2.7	0.310	
Total	11	21.4		

10.19 A company wants to compare a new brand of tennis ball with the old brand. The numbers below give the speed of the balls both old & new in mph.

Old brand	New brand
140	150
150	165
135	170
180	180
190	195
160	200

Compare the two balls and find out if there are any significant differences among them in terms of their speed. Carry out the test.

- Using the methods given in chapter 7.
- Analysis of variance (using 5% level of significance)

10.20 The figures given below give the nicotine content in a pack of cigarettes for 4 brands.

A	B	C	D
290	270	250	230
230	260	245	240
243	250	270	210
	280		200

At 5% level of significance, test if there are any significant differences among the brands.

- 10.21 The performance of business school students in 4 different subjects of this course gave the following scores:

Marketing	HR	QT	Finance
70	65	80	85
75	70	85	82
60	75	87	86
78	77	81	88

Is there any significant difference in the performance of the students across the four subjects?

- 10.22 A laboratory tested 3 bulbs manufactured by different companies to determine their performance in terms of their lumen. The test gave the following results:

Manufacturer A	Manufacturer B	Manufacturer C
845	875	790
800	880	795
875	890	798
825	888	799

Is there any significant difference in the performance of the bulbs of the three companies?
(Use 5% level of significance)

- 10.23 Four workers are working on 4 machines in a factory doing the same operation which involves testing a product. The management wants to test if the productivity of all four workers is same or there is some variance in their productivity. For this, data on the number of products each worker completes in a day is taken up randomly over a period of one month. This information is given in the following table:

Worker 1	Worker 2	Worker 3	Worker 4
550	700	450	750
680	720	300	600
600	750	520	650
650	780	250	680

Find out if the productivity of all four workers is same or different. (Use $\alpha = 5\%$ level of Significance).

- 10.24 A pharmaceutical company is testing 3 new drugs developed to cure cold. The three drugs were tested on patients to test if there are any differences in their effectiveness. The following data gives the no. of people who reported significant recovery in hospital located in three different cities. Find if there is any difference in their effectiveness at 5% level of significance.

- 10.25 An organization selling four products wants to determine whether the sales are distributed similarly among four general categories of customers. A random sample of 60 sales records provides the following information:

Customer Group	Product Type			
	1	2	3	4
Working executives	17	30	40	48
Businessmen	30	40	50	20
Academician	70	50	45	50
Others	40	20	37	34

Test at 90 percent confidence level, if the sales of the four products are independent of the customer groups.

- 10.26 Random samples of 160 persons from Mumbai, 240 persons from Delhi, 200 persons from Chennai were asked as to what kind of programmes on television did they prefer to watch. The responses are summarized in the following table:

Type of Programme	Number of persons		
	Mumbai	Delhi	Chennai
Real Life Show	60	100	80
Musical Programme	30	30	30
Serials	30	40	50
Talk Shows	40	70	40

At 95% confidence level, test if there is any significant difference in the proportion of people from the three cities preferring types of programmes.

- 10.27 The following data pertain to the number of units of a product manufactured per day by five workmen from four different brands of machines.

Workmen	Machine Brands			
	A	B	C	D
1	46	40	49	38
2	48	42	54	45
3	36	38	46	34
4	35	40	48	35
5	40	44	51	41

- (i) Test whether the mean productivity is the same for the four brands of machine type.
 (ii) Test, whether five different workmen differ with respect to productivity.

(M.Com., D.U, 1999)

10.28 There are three brands of a certain powder. A set of 120 sales is examined and found to be allocated among four groups (A, B, C, D) and brands (I, II, III) as shown below:

		Replications				
		Brands	A	B	C	D
Factor	I	0	4	8	15	
	II	5	8	13	6	
	III	18	19	11	13	

Check whether the factor “Brand” has significant effect on the sales at $\alpha = 0.05$ using one-way ANOVA.

(MBA, Bharathidasan Univ., April 2001)

10.29 The R & D manager of an automobile company wishes to study the effect of “Tyre Brand” on the tread loss (in millimetre) of tyres. Four tyres from each of four different brands (A, B, C, D) are fitted to four different cars using the completely randomized design. The data as per this design are presented below:

Tyre Brand				
	A	B	C	D
	6	3	8	4
	7	6	6	2
	10	2	7	1
	9	3	2	4

- (i) Write the corresponding model.
 (ii) Check whether the tyre brand has effect on the tread loss of tyres at a significant level of 5%.

(MBA, Bharathidasan Univ., 2002)

10.30 As head of a 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 lifetime of three electric bulbs of each of the four brands. The data is shown below, each entry representing the lifetime of an electric bulb, measured in hundreds of hours:

Brand			
A	B	C	D
20	25	24	23
19	23	20	20
21	21	22	20

Can we infer that the mean lifetime of the four brands of electric bulbs are equal?

(MBA, Univ., of Rorkee, 2000)

- 10.31 Since price and quality are supposed to be related, a consumer's forum decided to study the consumer preferences of about 100 products of daily general use. The following data was obtained:

Price	Bad	OK	Good
Less	10	15	5
Medium	8	20	12
Expenses	7	10	13

Test whether price and quality are related.

- 10.32 100 employees are working as computer data entry operators in a large electronics company. They appeared in an examination to be eligible for a promotion. The marks obtained by the 100 employees are as follows:

Marks	30-40	40 - 50	50 - 60	60 - 70	70 - 80
No. of Employees	10	20	30	25	15

Test if the distribution of marks can be approximated by a normal distribution.



11

Non-Parametric Tests



Structure

- 11.1 Introduction
- 11.2 One-Sample Runs Test
 - 11.2.1 Runs Test for Small Samples
 - 11.2.2 Runs Test for Large Samples
- 11.3 The Sign Test for Paired Observations
- 11.4 Rank Sum Tests
 - 11.4.1 Mann-Whitney U-test
 - 11.4.2 The Kruskal-Wallis H-Test
- 11.5 The Kolmogorov-Smirnov Goodness-of-fit Test
- 11.6 Exercises

11.1 INTRODUCTION

In chapter 7, we discussed various tests related to population parameters, both for a single sample and for two samples. While discussing these tests, an underlying assumption was that the parent population follows a normal distribution. These tests come under the purview of parametric tests or classical tests. In real life, there may be situations where such an assumption regarding normality of the population may not hold. The tests, which do not make any restrictive assumptions about the population parameters, are called non-parametric tests. In this chapter, we will discuss few common and widely used non-parametric tests.

The advantage of these tests are that they are easy to calculate and simple to understand in terms of applicability, than the classical tests. Also, they can be used in case the data is qualitative in nature and classical tests cannot be applied.

The downside of these tests is that most of these tests only consider the relative ranking of any observation and not its actual numerical value. Thus, a lot of information is wasted. Further, if the assumptions of parametric tests hold, then non-parametric tests are less efficient than parametric tests.

11.2 ONE - SAMPLE RUNS TEST

In most of our discussions related to various concepts of statistics, we have often considered 'random samples'. The random nature of the sample was accepted without any questions. However, an important question is that of testing if a sample is *actually random or non-random*. The runs test is a test for testing the randomness of a sample.

As the name suggests, this test is based on the concept of a run. We first define a run with an example.

Suppose the number of males (M) and females (F) coming for a new movie show is as follows:

MM FF MMM FF M FFF MMMM

Now in a sequence of letters of two kinds as above, a run is defined as a continuous and uninterrupted occurrence of letters of one kind. In the above sequence, first two M's occur. This is a run of 2 M's. Next two females are recorded. This is a run of 2 females and so on. The total number of runs in this sequence is 7.

11.2.1 Runs Test For Small Samples

Let m = the number of symbols of one kind.

n = the number of symbols of other kind.

R = the number of runs.

If m and n is less than or equal to 10 it is considered a small sample. For randomness, both symbols are expected to occur in equal numbers. Therefore, the null and the alternative hypothesis are:

H_0 : The occurrence of runs is random.

H_1 : The occurrence of runs is non - random.

For any combination of m and n , two critical values exist, at a level of significance 0.05, an upper critical value and a lower critical value. (Value available in tables) If the observed number

of runs (R) lies between the lower and the upper critical value accept the null hypothesis, else reject the null hypothesis. (The lower and upper critical values are available in tables)

Example 11.1: The following ups (u) and downs (d) were recorded in gold prices over 10 days.

u d d u u u d d u d

Test if this sequence is random.

Solution:

The runs are

u dd uuu dd u d

1 2 3 2 1 1

Thus there are 6 runs, 3 runs for ups and 3 runs for down.

Thus

$$m = 5$$

$$n = 5$$

The total number of runs = 6

The null hypothesis

H_0 : The movement of the gold prices is random.

The alternative hypothesis

H_1 : The movement of the gold prices is not random.

From tables, the critical values of r for $m = 5$ and $n = 5$ are 2 and 10 respectively at 5% level of significance.

Decision

Since the total number of runs is 6 in this case and it lies between 2 and 10 we may accept the null hypothesis.

Conclusion

The movement of the gold prices seems to be random.

Example 11.2: Consider 10 tosses of a coin. The results are recorded as below:

H T H T T H H H T H

Test at 5% level of significance if the results are random and the coin is unbiased.

Solution:

The null hypothesis

H_0 : the coin is unbiased.

The alternative hypothesis

H_1 : The coin is biased.

The runs are

H T H TT HHH T H

Total number of runs, $r = 7$

Thus

$$m = 6$$

$$n = 4$$

From tables, the critical values of r for $m = 6$ and $n = 4$ are 2 and 9 respectively for 5% level of significance.

Decision

Since the total number of runs is 7 and it lies between 2 and 9 respectively, we may accept the null hypothesis.

Conclusion

The coin seems to be unbiased.

11.2.2 Runs Tests for Large Samples

If both m and n are more than 10, than the sample is considered to be a large sample. And in this case, the distribution of R – the number of runs is approximately normal with mean

$$\mu_R = \frac{2mn}{m+n} + 1$$

and standard deviation

$$\sigma_R = \sqrt{\frac{2mn(2mn - m - n)}{(m+n)^2(m+n-1)}}$$

By transforming to the standard normal variate

$$Z = \frac{R - \mu_R}{\sigma_R}$$

where the distribution of Z is standard normal. The null and the alternative hypothesis remain same as the runs test for small samples. Since the alternative hypothesis is two tailed the acceptance region lies between $-Z_{\frac{\alpha}{2}}$ and $Z_{\frac{\alpha}{2}}$. If calculated Z lies in the acceptance region accept the null hypothesis, else reject it at α % level of significance.

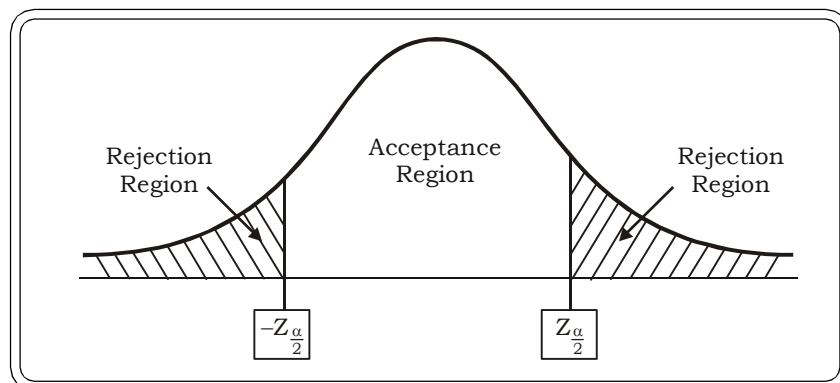


Figure 11.1

Standard Normal Curve

Example 11.3: An economist was interested in analyzing a stock market index to see if the movement of the index is random or it has some pattern. He recorded the observations continuously for a month giving a up (U) if it moves up and down (D) if the index shows a downward movement. The 30 observations recorded by him are:

DDUUDDDUDDDDUUUDDDDUUUDDDUUDU

Check if the movement of the index is random at 5% level of significance.

Solution:

The null hypothesis

H_0 : The movement of the index is random.

The alternative hypothesis

H_1 : The movement of the index is non – random.

Let m = number of D's
 n = number of U's
 R = number of runs

Then

$$\begin{aligned} m &= 17 \\ n &= 13 \\ R &= 12 \end{aligned}$$

The test statistic

$$Z = \frac{R - \mu_R}{\sigma_R}$$

where

$$\begin{aligned} \mu_R &= \frac{2mn}{m+n} + 1 \\ &= \frac{2(17)(13)}{17+13} + 1 \\ &= \frac{442}{30} + 1 = 15.73 \end{aligned}$$

$$\begin{aligned} \sigma_R &= \sqrt{\frac{2mn(2mn - m - n)}{(m+n)^2(m+n-1)}} \\ &= \sqrt{\frac{2(17)(13) \{2(17)(13) - 17 - 13\}}{(30)^2(17+13-1)}} \\ &= \sqrt{\frac{(442)(412)}{(900)(29)}} \end{aligned}$$

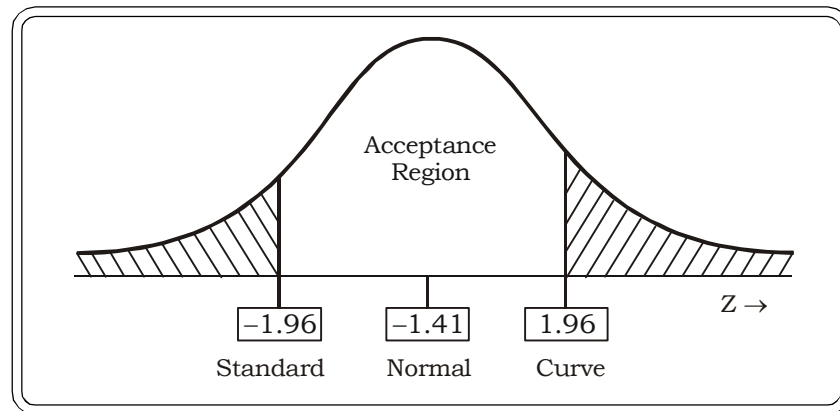
$$= \sqrt{\frac{182104}{26100}} = 2.64$$

Thus

$$Z = \frac{12 - 15.73}{2.64}$$

$$= -1.41$$

$$Z_{\frac{0.05}{2}} = Z_{0.025} = 1.96 \text{ (Since the test is two tailed)}$$



Decision

The calculated value of Z lies in the Acceptance Region.

Therefore, the null hypothesis may be accepted.

Conclusion

The economist may conclude that the index shows a random movement.

Example 11.4: A marketing manager is interested in determining if males and females visit a mall randomly or some kind of pattern exists. For example females might prefer to shop during the day and so on. He collected data from a mall by recording 40 observations and this is the pattern he got:

FF MMM FFF M FF MM FF M FF MM

MM FF MMM F MM MM FFF MFF MF

Test at 5% level of significance if this pattern is random.

Solution:

H_0 : The pattern of males and females visiting malls is random

H_1 : The pattern of males and females visiting malls is non – random

R = number of runs = 20

Let m = number of females

n = number of males

Then

$$m = 20$$

$$n = 20$$

$$\text{mean: } \mu_R = \frac{2(20)(20)}{20 + 20} + 1$$

$$= \frac{800}{40} + 1$$

$$= 21$$

$$\text{Standard Deviation: } \sigma_R = \sqrt{\frac{2(20)(20) \{2(20)(20) - 20 - 20\}}{(40)^2 (20 + 20 - 1)}}$$

$$= \sqrt{\frac{(800)(760)}{(1600)(39)}}$$

$$= \sqrt{\frac{608000}{62400}} = 3.12$$

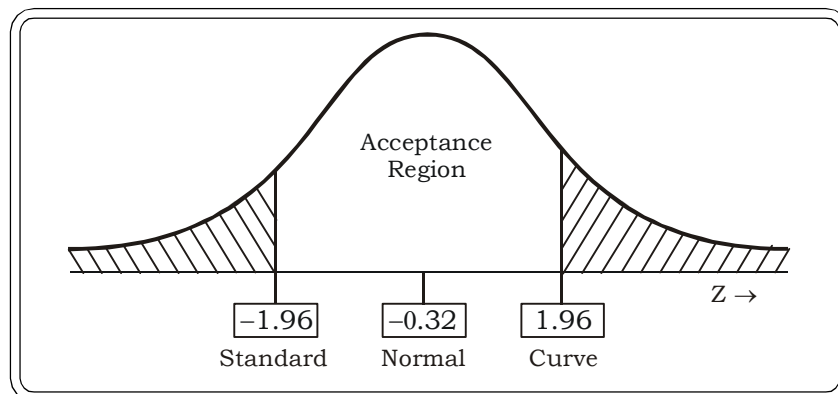
The test statistic

$$Z = \frac{R - \mu_R}{\sigma_R}$$

$$= \frac{20 - 21}{3.12} = -0.32$$

$$Z_{0.025} = 1.96$$

The test is two tailed.



Decision

The null hypothesis is accepted at 5% level of significance.

Conclusion

The arrival of men and women in malls is a random event.

11.3 THE SIGN TEST FOR PAIRED OBSERVATIONS

This test, also known as the two – sample sign test may be considered as the non – parametric equivalent of the paired t – test for dependent samples. As in the case of the paired t – test, the sign test is also used to test paired observations in two samples that are dependent in nature. Typical applications of the sign test for paired observations are:

- (i) To test the employee satisfaction levels before and after a significant raise in salaries of employees of the company.
- (ii) The number of hours required to manufacture a product before and after the machine was overhanded in a certain factory.
- (iii) To test whether a new diet to be introduced in the market is effective in reducing the weights of individuals. The two samples in this case will consist of weights of individuals before the diet was administered and the weights of the individuals after the diet was administered.

The sign test does not consider the actual numerical values of the data. Rather, it considers the plus and minus signs as the basis for testing. Let us consider the first example as an illustration. In order to test whether employee satisfaction levels have gone up post a significant salary raise and certain new employee friendly policies introduced by the management, the following data was obtained (on a scale of 100):

Table 11.1

Employee Satisfaction Levels Before and After Salary Raise

Before	After
72	79
74	82
81	75
83	87
86	97
98	94
60	70
79	71
72	80
83	87

We now compute the sign of the difference between the two sets of observations. A plus sign (+) if the first component exceeds the second and a minus sign (–) otherwise

Table 11.2
Table of Signs

Before	After	Sign
72	79	-
74	82	-
81	75	+
83	87	-
86	97	-
98	94	+
60	70	-
79	71	+
72	80	-
83	87	+

It must be noted here that in case of ties the pair for which a tie occurs is eliminated from the sample and subsequently the sample size n also gets reduced. In this example there are no ties and the sample size is $n = 10$.

The null hypothesis and the alternative hypothesis

H_0 : There is no difference in the weights of the individuals before and after the diet was administered.

H_1 : There is significant difference in the weights of the individuals before and after the diet was administered.

If the null hypothesis were true, one would expect an equal number of plus and minus signs.

Let

X : the number of plus signs.

p : probability of getting a plus sign.

The distribution of X can be approximated by a Binomial distribution with $p = \frac{1}{2}$.

Thus the null hypothesis being tested is

$$H_0: p = \frac{1}{2}$$

Against the alternative hypothesis

$$H_1: p \neq \frac{1}{2}$$

Here $N = 10$

The number of plus sign (X) = 4

The number of minus sign = 6

This is similar to the parametric test for a single proportion. Now, under the following conditions we can use a normal approximation to the binomial distribution:

- (i) $Np \geq 5$
- (ii) $Nq \geq 5$
- (iii) p is not close to 0 or 1

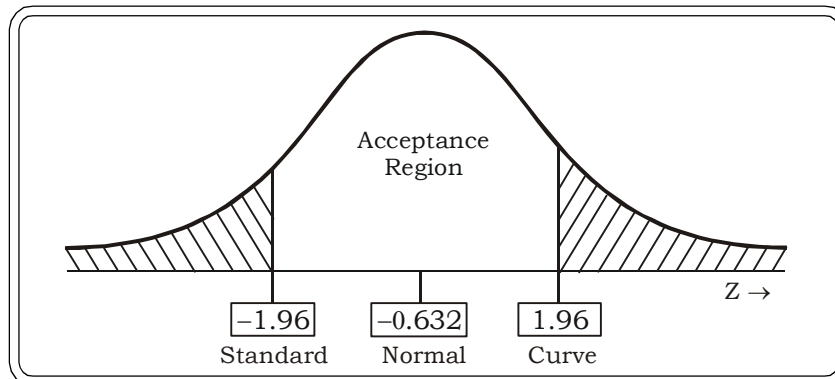
In this case all three conditions are satisfied.

The test statistic

$$Z = \frac{\frac{X}{N} - \frac{1}{2}}{\sqrt{\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)/N}}$$

$$= \frac{\frac{4}{10} - \frac{1}{2}}{\sqrt{\left(\frac{1}{4}\right)/10}} = \frac{0.4 - 0.5}{0.158}$$

$$= -0.632$$



Since the critical value falls in the critical region, we may accept the null hypothesis.

Conclusion

Thus there is no evidence to indicate that there is significant difference in the weights of the individuals before and after administration of the diet.

Alternatively, by computing the mean and the standard deviation as

$$\mu = Np \text{ and } \sigma = \sqrt{Npq}$$

We can also use the test statistic

$$Z = \frac{X - \mu}{\sigma}$$

to test the null hypothesis.

Here

$$\mu = 10 \frac{1}{2} = 5$$

$$\sigma = \sqrt{10 \times \frac{1}{2} \times \frac{1}{2}} = 1.58$$

Thus

$$Z = \frac{4 - 5}{1.58} = -0.63$$

which is same as the value of Z previously obtained.

Example 11.5: In a bid to improve the performance of sales executives of a company, the management decided to administer a one-month course in sales and marketing to 15 of its salesmen. The sales of these executives before and after the course were recorded in order to evaluate the effectiveness of the course. Test, at 5% level of significance whether the course has proved beneficial in improving the performance of the sales executives of the company.

Salesman	Sales Before Course	Sales after Course
1	11	13
2	10	12
3	12	9
4	8	8
5	7	9
6	6	10
7	12	10
8	6	12
9	9	8
10	8	7
11	7	8
12	5	5
13	10	12
14	10	9
15	12	10

Solution:

We first need to compute the signs of difference:

Salesman	Sales Before Course	Sales after Course	Sign
1	11	13	-
2	10	12	-
3	12	9	+
4	8	8	Tie
5	7	9	-
6	6	10	-
7	12	10	+
8	6	12	-
9	9	8	+
10	8	7	+
11	7	8	-
12	5	5	Tie
13	10	12	-
14	10	9	+
15	12	10	+

X: the number of plus signs

Here

$$X = 6$$

N: the sample size

$$N = 13 \text{ (since there are two ties)}$$

The null hypothesis

H_0 : There is no difference in the sales before and after the course i.e. $p = \frac{1}{2}$

The alternative hypothesis

H_1 : There is significant difference in the sales before and after the course was administered on the executives i.e. $p \neq \frac{1}{2}$

To apply the normal approximation, we first calculate the mean and the variance as follows:

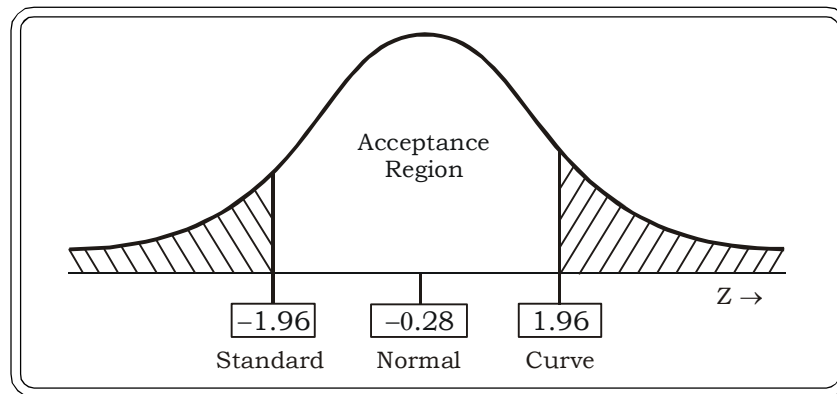
$$\mu = Np = \frac{13}{2} = 6.5$$

$$\sigma = \sqrt{Npq} = \sqrt{\frac{13}{4}} = 1.80$$

The test statistic

$$Z = \frac{6 - 6.5}{1.80} = -0.28$$

The Z value at 5% level of significance for a two tailed test = ± 1.96 .



Decision:

Thus the calculated value of Z lies in the acceptance region. We accept the null hypothesis.

Conclusion

There is no significant difference in the performance of the sales executives after taking the one - month course, indicating that perhaps the course was not very effective.

11.4 RANK - SUM TESTS

The rank sum tests are used to test whether two or more samples drawn independently come from identical populations. For two samples, the Mann - Whitney U test is used and for more than two samples the Kruskal - Wallis H - test can be applied.

11.4.1 Mann - Whitney U test

In the t - test for testing the equality of two means, one assumption was that the parent populations must be normally distributed with equal variances. When this assumption cannot be met, a Mann - Whitney U test can be used in place of the classical t - test.

In this test, data from both the samples are combined into one and the observations are then arranged in ascending order. The smallest observation is ranked 1 and the next smallest is ranked 2 and so on till the highest gets the largest rank. In case of ties the mean rank of all the tied observations is calculated and then assigned to all the tied observations.

The null and the alternative hypothesis

H_0 : the two samples drawn independently, are from identical populations.

H_1 : The two samples drawn independently, are not from identical populations.

The Test Statistic

Let m be the size of one of the samples.

Let R_1 be the sum of the ranks assigned to observations in this sample

Let n be the size of the other sample,

then the test statistic is defined as:

$$u = mn + \frac{m(m+1)}{2} - R_1$$

Under the null hypothesis, the distribution of u has mean and standard deviation:

$$\mu = \frac{mn}{2} \text{ and}$$

$$\sigma = \sqrt{\frac{mn(m+n+1)}{12}}$$

If m and n are greater than 8, u can be approximated by a normal distribution.

Standardizing Z ,

$$z = \frac{u - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \sim N(0, 1)$$

If the calculated value of $|z|$ is less than tabulated value of $\frac{Z_\alpha}{2}$, we may accept the null hypothesis, else reject it at $\alpha\%$ level of significance.

We consider the following example to explain the application of this test.

Example 11.6: Consider the following data related to ages of people who liked a new music album and ages of people who did not like the new music album launched in the market.

Liked the album (L)	Did not like the album (D)
28	31
33	42
26	46
30	38
29	40
24	41
55	49
31	37
25	30
40	42
	35

Solution:

To find out whether the age distribution of people who liked the album and those who did not like the album is identical, we can now use the Mann –Whitney u test. A close examination of the two samples shows more variability in the first sample than in second sample. This clearly indicates a violation of the equality of variance assumptions.

To apply the rank sum test, we first combine the data and arrange the observations in ascending order and then assign ranks to them.

Ages	Sample	Rank	
24	L	1	
25	L	2	
26	L	3	
28	L	4	
29	L	5	
30	L	6.5	{ Mean of 6 & 7 }
30	D	6.5	
31	L	8.5	{ Mean of 8 & 9 }
31	D	8.5	
33	L	10	
35	D	11	
37	D	12	
38	D	13	
40	L	14.5	{ Mean of 14 & 15 }
40	D	14.5	
41	D	16	
42	L	17.5	{ Mean of 18 & 19 }
42	D	17.5	
46	D	19	
49	D	20	
55	L	21	

Ranks of people who liked the album are:

1, 2, 3, 4, 5, 6.5, 8.5, 10, 14.5, 17.5 and 21

Ranks of people who did not like the album:

6.5, 8.5, 11, 12, 13, 14.5, 16, 17.5, 19 and 20

Now, sum of ranks of the first sample is:

$$R_1 = 1 + 2 + 3 + 4 + 5 + 6.5 + 8.5 + 10 + 14.5 + 17.5 + 21 = 93$$

Sum of ranks of the second sample is

$$R_2 = 6.5 + 8.5 + 11 + 12 + 13 + 14.5 + 16 + 17.5 + 19 + 20 = 138$$

The null hypothesis and the alternative hypothesis

H_0 : The age distribution of people who liked the new music album and those who did not like the new music album are identical.

H_1 : The age distribution of people who liked the new music album and who did not like the new music album are different.

Let m = size of the first sample

$$= 10$$

and n = size of the second sample.

$$= 11$$

The test statistic

The test is now based on the statistic defined as

$$u = mn + \frac{m(m+1)}{2} - R_1$$

The statistic has mean

$$\mu = \frac{mn}{2}$$

and standard deviation

$$\sigma = \sqrt{\frac{mn(m+n+1)}{12}}$$

where m and n are both greater than 8, the distribution of u is normal. Thus, by converting to the Z scale the test statistic for testing H_0 is

$$Z = \frac{u - \mu}{\sigma}$$

In this example

$$m = 10, n = 11, R_1 = 93$$

$$\text{Therefore, } u = 10 \times 11 + \frac{10(10+1)}{2} - 93 = 110 + 55 - 93 = 72$$

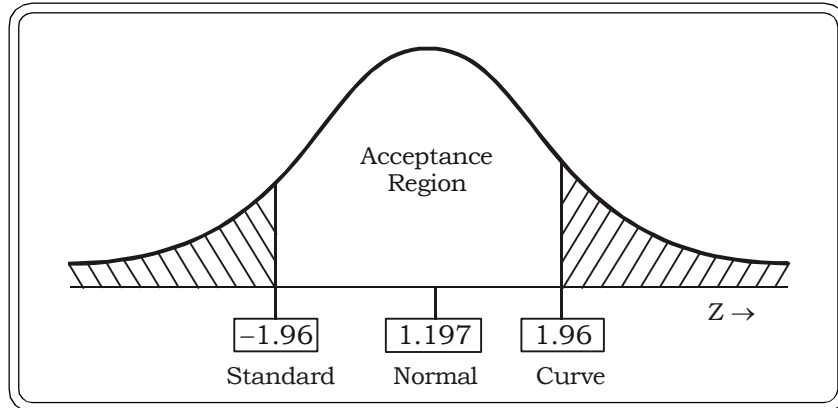
$$\mu = \frac{110}{2} = 55$$

$$\sigma = \sqrt{\frac{110(22)}{12}} = 14.20$$

Thus the test statistic is

$$Z = \frac{72 - 55}{14.20} = 1.197$$

At 5% level of significance the table values of $Z = 1.96$ (since the test is two tailed).



Decision:

Since the calculated value of Z lies in the acceptance region, we may accept the null hypothesis.

Conclusion:

There is no difference in the age distribution of people who liked the new music album and those who did not like the album. The music seems to appeal to people of all ages.

Example 11.7: A large store wants to compare if their customers preference to pay by credit card and cash is same. They collect the following data of sales by credit card and sales by cash in an effort to compare the amount of rupees paid by each. (Amount in 000's)

Credit Card (cc)	Cash (C)
90	80
100	105
80	51
120	85
105	110
102	95
99	87
95	96
85	

At 5% level of significance, test the hypothesis that there is no difference in the amount of money paid by credit card and by cash.

Solution:

We first combine data from both samples into one sample by arranging them in ascending order as follows:

Payment (in 000 rupees)	Mode of Payment	Rank
51	C	1
80	C	2.5
80	CC	2.5
85	C	4.5
85	CC	4.5
87	C	6
90	CC	7
95	CC	8.5
95	C	8.5
96	C	10
99	CC	11
100	CC	12
102	CC	13
105	CC	14
106	C	15
110	C	16
120	CC	17

$$\begin{aligned}
 R_1 &= \text{sum of ranks of the first sample} \\
 &= 2.5 + 4.5 + 7 + 8.5 + 11 + 12 + 13 + 14 + 17 \\
 &= 89.5
 \end{aligned}$$

$$\begin{aligned}
 R_2 &= \text{sum of ranks of the second sample} \\
 &= 1 + 2.5 + 4.5 + 6 + 8.5 + 10 + 15 + 16 \\
 &= 63.5
 \end{aligned}$$

The null hypothesis

H_0 : There is no significant difference in the amount of rupees paid by customers by credit card or cash.

The alternative hypothesis:

H_1 : There is significant difference in the amount of rupees paid by customers by credit card and by cash

Now

$m = 9$ (Number of people paying by Credit Card)

$n = 8$ (Number of people paying by Cash)

Thus

$$\begin{aligned} u &= mn + \frac{m(m+1)}{2} - R_1 \\ &= 72 + \frac{9 \times 10}{2} - 89.5 \\ &= 27.5 \end{aligned}$$

This statistic u has

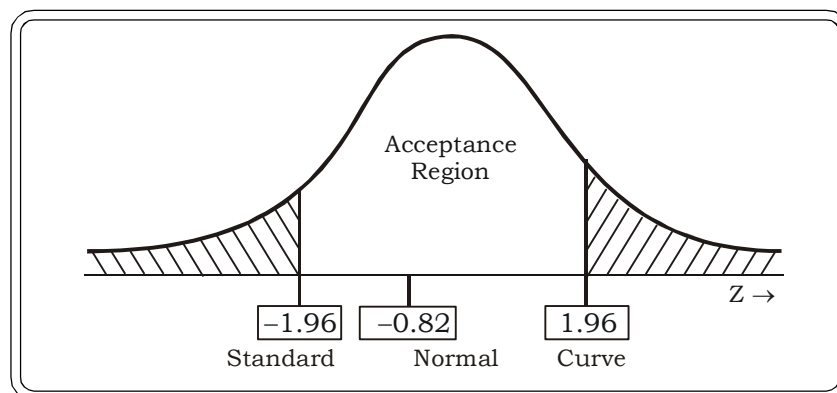
$$\text{Mean} = \frac{mn}{2} = 36$$

$$\begin{aligned} \text{Variance} &= \sqrt{\frac{mn(m+n+1)}{12}} \\ &= \sqrt{\frac{72(18)}{12}} = 10.39 \end{aligned}$$

The test statistic

$$\begin{aligned} Z &= \frac{u - \text{mean}}{\text{variance}} \\ &= \frac{27.5 - 36}{10.39} = -0.82 \end{aligned}$$

Since the test is two tailed, the tabulated value of $Z = \pm 1.96$



Decision:

Since the calculated value of z lies in the acceptance region we may accept the null hypothesis.

Conclusion:

At 5% level of significance, we may conclude that there is no difference in the amount of money paid by customers by cash or by credit card. Customers seem to prefer both modes equally.

11.4.2 The Kruskal – Wallis H – Test

This test is an extension of the Mann Whitney u – test. It is used to test if more than two independent samples are drawn from identical populations or not. In a way, this test may be considered as the non-parametric equivalent of one-way Analysis of Variance (ANOVA), discussed in chapter 10. The methodology is similar to the Mann Whitney U test. All observations are combined to form one sample. They are then ranked from smallest to the highest. Suppose there are p samples. Then the statistic H is defined as:

$$H = \frac{12}{N(N+1)} \left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \dots + \frac{R_p^2}{n_p} \right) - 3(N+1)$$

where n_1 = size of the first sample

n_2 = size of the second sample

.

.

n_p = size of the p^{th} sample

R_1 = sum of ranks in the first sample

R_2 = sum of ranks in the second sample

.

.

R_p = sum of ranks in p^{th} sample

$N = n_1 + \dots + n_p$

= size of the pooled sample

The null hypothesis:

H_0 : all the p independent samples come from identical populations.

The alternative hypothesis

H_1 : The p independent samples come from different populations.

Now, under the null hypothesis, provided that each sample size is at least 5, the sampling distribution of H can be approximated by a chi-square (χ^2) distribution with $(p - 1)$ degrees of freedom. The decision rule is to reject the null hypothesis if $\text{cal } \chi^2 > \text{tab } \chi^2$ with $(p - 1)$ degrees of freedom at a certain given level of significance.

Example 11.8: The following table is related to sales of 3 brands of shampoo in a store. The manager of the store would like to evaluate if the sales of all the 3 brands are same in his store. Use a Kruskal Wallis H-test to evaluate whether at 5% level of significance, the sales of the 3 brands of shampoo are same or different. Data related to sales of the three shampoo brands are:

Brand A	Brand B	Brand C
323	318	360
338	330	335
358	320	319
350	360	350
335	340	340
320	337	

Solution:

We first combine the observations in ascending order and rank them as follows:

Sales	Brand	Rank
318	B	1
319	C	2
320	B	3.5
320	A	3.5
323	A	5
330	B	6
335	C	7.5
335	A	7.5
337	B	9
338	A	10
340	B	11.5
340	C	11.5
350	A	13.5
350	C	13.5
358	A	15
360	B	16
360	C	17

The null hypothesis:

H_0 : there is no significant difference in the sales of the three brands of shampoo.

The alternative hypothesis:

H_1 : There is significant difference in the sales of the three brands of shampoo.

$$p = 3$$

$$n_1 = 6$$

$$n_2 = 6$$

$$n_3 = 5$$

$$N = 17$$

$$\begin{aligned} R_1 &= \text{sum of ranks of Brand A} \\ &= 3.5 + 5 + 7.5 + 10 + 13.5 + 15 \\ &= 54.5 \end{aligned}$$

$$\begin{aligned} R_2 &= \text{sum of ranks of Brand B} \\ &= 1 + 3.5 + 6 + 9 + 11.5 + 16 \\ &= 47 \end{aligned}$$

$$\begin{aligned} R_3 &= \text{sum of ranks of Brand C} \\ &= 2 + 7.5 + 11.5 + 13.5 + 17 \\ &= 51.5 \end{aligned}$$

The test statistic

$$\begin{aligned} H &= \frac{12}{N(N+1)} \left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} \right) - 3(N+1) \\ &= \frac{12}{306} (495.04 + 368.17 + 530.45) - 54 \\ &= 0.65 \end{aligned}$$

The calculated value of χ^2 with 2 degrees of freedom at 5% level of significance is 5.991.

Decision:

We may accept the null hypothesis since the calculated value of χ^2 is less than the tabulated value of χ^2 .

Conclusion:

There seems to be no significant difference in the sales of the three brands of shampoo in the store. All three brands seem to be selling equally well.

Example 11.9: Three different brands of cigarettes were tested for the tar content (in milligrams) in pack of cigarettes. Six packs of each brand were tested and the figures are given below:

Brand A	Brand B	Brand C
13	14	12
12	10	14
11	13	13
10	16	10
14	11	11
12	12	15

At 5% level of significance do the data provide evidence that the three brands differ significantly in terms of their tar content.

Solution:

The null hypothesis

H_0 : The three brands do not differ significantly in terms of their tar content.

The alternative hypothesis

H_1 : The three brands differ significantly in terms of their tar content.

Next, we pool all the three samples into one sample, arrange them in ascending order and rank them from the smallest to the highest.

Tea Content (in milligram)	Brand	Rank
10	A	2
10	B	2
10	C	2
11	A	5
11	B	5
11	C	5
12	A	8.5
12	A	8.5
12	B	8.5
12	C	8.5
13	A	12
13	B	12
13	C	12
14	A	15
14	B	15
14	C	15
15	C	17
16	B	18

$$N = 18$$

$$n_1 = 6, n_2 = 6, n_3 = 6$$

$$\begin{aligned} R_1 &= \text{sum of ranks of Brand A} \\ &= 2 + 5 + 8.5 + 8.5 + 12 + 15 \\ &= 51 \end{aligned}$$

$$\begin{aligned} R_2 &= \text{sum of ranks of Brand B} \\ &= 2 + 5 + 8.5 + 12 + 15 + 18 \\ &= 60.5 \end{aligned}$$

$$\begin{aligned} R_3 &= \text{sum of ranks of Brand C} \\ &= 2 + 5 + 8.5 + 12 + 15 + 17 \\ &= 59.5 \end{aligned}$$

The test statistic

$$\begin{aligned}
 H &= \frac{12}{N(N+1)} \left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} \right) - 3(N+1) \\
 &= \frac{12}{(18)(19)} \left(\frac{51^2}{6} + \frac{60.5^2}{6} + \frac{59.5^2}{6} \right) - 3(19) \\
 &= \frac{12}{342} (433.5 + 610.04 + 590.04) - 57 \\
 &= 0.32
 \end{aligned}$$

The tabulated value of χ^2 at 2 degrees of freedom for 5% level of significance is 5.991

Decision

Since the calculated H is less than tabulated χ^2 we may accept the null hypothesis.

Conclusion

At 5% level of significance there is no reason to doubt that the tar content of the three brands of cigarettes are significantly different.

11.5 THE KOLMOGOROV SMIRNOV GOODNESS - OF - FIT TEST



The Kolmogorov-Smirnov test is similar to the χ^2 test of goodness-of-fit in terms of applicability as it is also used to test distributional adequacy of a sample. This test is used to test whether there is any significant difference between an observed frequency distribution and an expected frequency distribution. This test is based on the difference between observed relative cumulative frequencies. If this difference is small, then this would indicate not much difference between the two distributions. However a significantly large difference between any two points could indicate that the two distributions are different. This is the basic premise on which this test is based.

The test statistic

The Kolmogorov - Smirnov statistic also known as the 'D - statistic' is defined as

$$D_n = \max |CF_e - CF_o|$$

Where

CF_e = expected cumulative relative frequency

CF_o = observed cumulative relative frequency

The critical values of D_n are given in tables.

The decision rule is to accept the null hypothesis if the cal D_n is less than the table value of D_n and reject otherwise.

Let us now consider an example of application of the Kolmogorov - Smirnov goodness of fit test.

Example 11.10: The distribution of the number of clients a chartered accountant had on each day, during 400 randomly chosen days is as follows:

No. of Clients	No. of Days
0	50
1	112
2	125
3	60
4	28
5	16
6	9

The theoretical or expected number of clients is as follows:

No. of Clients	No. of Days
0	40
1	108
2	108
3	72
4	36
5	20
6	16

Test at 5% level of significance if there is any difference between the observed and the expected frequencies of the above data by using the Kolmogorov – Smirnov goodness – of – fit test.

Solution:

The null hypothesis

H_0 : There is no significant difference between the observed and the expected frequencies.

The alternative hypothesis

H_1 : There is significant difference between the observed and the expected frequencies.

To apply the Kolmogorov – Smirnov test, we first have to convert the frequencies into relative cumulative frequencies as given in the following table:

No. of Clients	Observed Frequencies	Expected Frequencies	Observed Cumulative frequency	Expected Cumulative Frequency	CF _o	CF _e
0	50	40	50	40	0.125	0.1
1	112	108	162	148	0.405	0.37
2	125	108	287	256	0.718	0.64
3	60	72	347	328	0.868	0.82
4	28	36	375	364	0.938	0.91
5	16	20	391	384	0.978	0.96
6	9	16	400	400	1.0	1.0

We now compute absolute differences between F_o & F_e

CF _o	CF _e	CF _e - CF _o
0.125	0.1	0.025
0.405	0.37	0.035
0.718	0.64	0.078
0.868	0.82	0.048
0.938	0.91	0.028
0.978	0.96	0.018
1.0	1.0	0.00

$$D_n = \max |CF_e - CF_o|$$

$$= 0.078$$

For a sample size of 400 the table value of D_n at 5% level of significance can be obtained by the formula

$$\frac{1.36}{\sqrt{n}} = \frac{1.36}{\sqrt{400}} = 0.068$$

Decision

Since the calculated value of D_n is more than the tabulated value of D_n. We reject the null hypothesis.

Conclusion

At 5% level of significance, there seems to be significant difference between the observed frequencies and the theoretical expected frequencies.

Example 11.11: The city police department is examining records of 50 busy intersections of the city, randomly selected to determine the number of accidents at these traffic intersections. The following data emerged from past records:

No. of Accidents	No. of Intersections
0	6
1	10
2	15
3	12
4	7

The head of the department (HOD) is of the opinion that this data can be reasonably approximated by a Poisson distribution with mean 2. At 5% level of significance test the opinion of the HOD.

Solution

The null hypothesis

H_0 : A Poisson distribution with $\lambda = 2$ is a good description of the given data.

The alternative hypothesis

H_1 : A Poisson distribution with $\lambda = 2$ is not a good approximation of the given data.

We have to now calculate the expected frequencies for a Poisson distribution with $\lambda = 2$ Using Poisson table the expected frequencies are calculated as follows:

No. of Accidents (X)	p(X)	Exp F (X) = Np(X)
0	0.1353	7
1	0.2707	13
2	0.2707	14
3	0.1804	9
4	0.1429	7
		50

The probabilities may also be calculated by using the recurrence relation given in section 5.8.3.6, Chapter 5.

To calculate the Kolmogorov – Smirnov statistic, we next compute the differences between the observed relative cumulative frequencies and the expected relative cumulative frequencies.

Observed Frequencies	Expected Frequencies	Observed Cumulative Frequency	Expected Cumulative frequency	CF _o	CF _e	CF _e - CF _o
6	7	6	7	0.12	0.14	0.02
10	13	16	20	0.32	0.4	0.08
15	14	31	34	0.62	0.68	0.06
12	9	43	43	0.86	0.86	0
7	7	50	50	1	1	0

$$D_n = \max |CF_e - CF_o| = 0.08$$

The table value of the D - statistic for a sample of size at 5% level of significance is

$$\frac{1.36}{\sqrt{50}} = 0.19$$

Decision

The calculated value of the D - statistic is smaller than the tabulated value. Therefore we may accept the null hypothesis.

Conclusion

At 5% level of significance, the opinion of the head of the city police department that the data related to accidents in busy traffic intersection can be approximated by a Poisson distribution with mean 2 is correct.

Table 11.3

Comparison of Classical Tests and their Non-Parametric Equivalent Test

Classical Test	Equivalent Non-Parametric Test
Paired t - test	The sign test for Paired observation.
t - Test for equality of two means	Mann - Whitney U - test
One way ANOVA	The Kruskall Wallis H Test
χ^2 test of goodness - of - fit	The Kolmogorov - Smirnov Goodness-of-fit test

11.6 EXERCISES



- 11.1 A professor conducted a test for his students consisting of 30 true (T) / False (F) questions. The sequence of answers for his test is as follows:

T F F T T T F T F T F F T F T T F F T F

F T T T F F T F T T

The Professor is concerned about whether or not the questions have been arranged in a random manner. Conduct a test for randomness of the arrangement of questions.

- 11.2 In a bulb factory, regular inspection is done to determine the defectives and non-defectives in a lot. On a certain day, a lot size of 40 gave the following data. (d – defective, g – good).

g b b b g g g g b g b g g g b b g g b g

g g g b g g b g b b g g b b g g g b g b

Test at 5% level of significance if the defectives are occurring randomly.

- 11.3 This is a sentence from a book – “As I sat on top of the foyer, my eyes went straight to the busy street in front. Hundreds of cars were driving fast. People were in a great hurry to move forward and appeared to have little time at hand.” In the above sentence, test whether the consonants and the vowels occur at random. Use 5% level of significance).

- 11.4 A financial analyst wants to check if the movement of the BSE Index is random in the past month. He collected data for 30 days and organized it as follows (U – upward movements, D – downward movements)

U U U D D U D D D U U D D U U U D U D D

D D U U U D U D U U

Use the runs test to determine whether the BSE Index movement is random or not.

- 11.5 It is believed that yoga and meditation helps immensely in bringing down blood pressure of individuals. A study was conducted to examine if there is any truth in this belief. 15 subjects were selected for the study. The systolic blood pressure of 15 individuals was recovered before and after the start of a one-hour yoga and meditation session. Test, at 5% level of significance, whether there is any dip in the blood pressure levels after the yoga and meditation session.

Subject No.	Systolic BP Before	Systolic BP After
1	130	130
2	135	132
3	140	135
4	150	155
5	125	130
6	145	140
7	170	165
8	155	150
9	150	150
10	175	170
11	160	150
12	140	140
13	147	145
14	150	145
15	138	137

11.6 A psychologist was studying the effects of supervision on workers in a factory shop floor. In particular he wished to determine if supervision has any effect on productivity of the workers. A sample of fifteen workers was selected for the study. The scores of productivity measured on the workers with supervision and without supervision in a study spread over two months are given in the table below:

Worker No.	Productivity Scores	
	With supervision	Without supervision
1	82	85
2	80	80
3	99	95
4	88	89
5	78	79
6	82	82
7	93	91
8	85	83
9	96	96
10	87	80
11	98	88
12	89	92
13	92	98
14	86	86
15	84	90

At 5% level of significance, determine if supervision affects productivity.

- 11.7 A pharmaceutical company has developed a new version of a drug to cure the common cold. It wants to conduct a test to check its effectiveness as compared to the older version of the drug. Out of 20 patients selected, 10 were given the old drug and the rest were given the new drug. The recovery time recorded for all the 20 patients are as follows:

Old version of drug (Recovery days)	New version of Drug (Recovery days)
7	5
8	4
9	6
6	7
8	6
7	6
5	5
5	6
6	7
7	5

Test, at 5% level of significance, using the Mann Whitney test, if there is any significant difference in the recovery times of the two drugs.

11.8 Test whether the following two data sets come from independent populations or from the same population.

Sample 1	Sample 2
27	29
22	10
31	33
15	38
40	15
32	24
27	10
31	12
20	40

11.9 Two machines A and B are producing the same product in a certain factory. The operator wants to study if both machines are taking the same amount of time to produce the product or different times. He records the time taken by each machine to produce 10 products as follows:

Machine A Time in hours	Machine B Time in Hours
4.2	4.7
4.5	4.8
4.0	4.5
3.9	5.0
4.7	3.9
4.6	4.4
4.4	4.6
4.2	4.9
4.3	4.5
4.1	4.3

Test, at 5% level of significance, whether the time taken by both machines are different or same.

- 11.10 A retail chain has four stores located at four different malls in four locations of the city. The management of the retail chain wants to find out if more customers come to certain stores or all the four stores have the same number of customers. A record of 10 randomly selected days gives the following data across all the four stores:

Store 1	Store 2	Store 3	Store 4
126	145	120	107
122	140	130	110
120	138	140	112
127	147	129	115
130	137	140	100
135	150	155	120
140	140	160	114
125	138	135	105
120	130	132	103
132	142	145	102

11.11 Test, using Kruskal – Wallis Test, if the following samples come from the same population or from different populations.

Sample 1	Sample 2	Sample 3
30	31	26
32	30	22
29	28	20
28	27	26
26	22	25
27	23	30
25	29	29
26	28	30

11.12 Three sections of the students of a university were taught finance by three different professors. The marks obtained by the students in a test are as follows:

Prof. A	Prof. B	Prof. C
60	59	62
62	58	63
65	60	64
70	62	65
69	61	62
68	59	61
61	58	60
63	57	59

Test, using the Kruskal – Wallis Test, if the marks obtained by the students taught by the three different professors are same or different.

11.13 In a call center, an executive recorded the following calls:

No. of Calls	No. of Days
0	22
1	46
2	65
3	55
4	50
5	45
6	27

Test, using the Kolmogorov – Smirnov goodness - of - fit test, if the data can be approximated by a Poisson distribution.

11.14 Test, using the Kolmogorov Smirnov Test, if the following data related to sales of a company can be approximated by a binomial distribution:

No. of sales per day	Frequency of no. of sales
0	15
1	55
2	39
3	22
4	12
5	8

11.15 The manager of the production department is concerned about the workers reporting late for duty. He collected data on 10 days and found the number of late reporting as:

3 5 2 1 5 2 1 1 6 3

Verify, using the sign test if the mean no. of workers reporting late for duty is more than 3.

